Universidad Autónoma
de Madrid

Escuela Politécnica Superior

Departamento de Tecnología Electrónica y de las Comunicaciones

COMBINING LOCAL FEATURES AND REGION SEGMENTATION: METHODS AND APPLICATIONS

PhD Thesis written by
Fulgencio Navarro Fajardo
under the supervision of
Dr. Jesús Bescós Cano
and
Dr. Marcos Escudero Viñolo

Madrid, November 2019

Department: Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid, Spain

PhD Thesis: Local features and region segmentation algorithms: characteristics segregation based on mutual analysis, alternatives and applications.

Author: **Fulgencio Navarro Fajardo**
Ingeniero de Telecomunicación
Universidad Autónoma de Madrid, Spain

Supervisor: **Jesús Bescós Cano**
Doctor Ingeniero de Telecomunicación
Universidad Autónoma de Madrid, Spain

Supervisor: **Marcos Escudero Viñolo**
Doctor Ingeniero de Telecomunicación
Universidad Autónoma de Madrid, Spain

Year: 2019

Comittee: **Fernando Marqués Acosta**
Universidad Politécnica de Cataluña, Spain

Jose Mª Martínez Sánchez
Universidad Autónoma de Madrid, Spain

Luis Salgado Álvarez de Sotomayor
Universidad Politécnica de Madrid, Spain

The work described in this Thesis was carried out within the Video Processing and Understanding Lab at the Department of Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid (from 2014 to 2019). It was partially supported by the Spanish Government (TEC2014-53176-R, HAVideo).

To my family and friends.

If opportunity doesn't knock, build a door.
- Milton Berle

Acknowledgments

I would like to express my gratitude to my supervisors, Dr. Jesús Bescós and Dr. Marcos Escudero Viñolo. I want to thank all my VPU collegues, Rafael Martín, Alvaro García, Jose Mª Martínez and the rest of the team. I also want to thank the people who have helped with the writing and validation procedures of the thesis, specially to the department reviewer Dr. Manuel Sánchez. Thanks to Dr. Fernando Marqués, Dr. Jose María Martínez and Dr. Luis Salgado for being part of the tribunal of this Thesis.

Thanks to my friends, specially to Juan Manuel, a brother, your support has been key for accomplishing such a project. Thanks to the people from Sigma Technologies and Tax Planning, I feel lucky to be part of such a great family.

Thanks to all my family. To my parents, you are responsible of all my past, present and future successes. To my brother, the most valuable and undeserved gift I have received in life.

Julia, my wife and the love of my life. You are the guiding light of this boat, please, remember to never stop shining.

To those not included here, you know very well that this would not have been possible without your help. Thank you all.

Fulgencio Navarro Fajardo
November 2019

Abstract

A huge number of proposals have been developed in the area of computer vision for information extraction from images, and its further use. One of the most prevalent solutions are those known as local features. They detect points or areas of the image with certain characteristics of interest, and describe them using information from their (local) environment. The regions also stand out in the area, and especially this work has focused on the region segmentation algorithms, whose objective is to group the information of the image according to different criteria.

Despite the enormous potential of these techniques, and their proven success in a number of applications, their definition implies a series of functional limitations that have prevented them from exporting their capabilities to other application areas. In this thesis, it is intended to promote the use of these tools in these applications, and therefore improve the results of the state of the art, by proposing a framework for developing new solutions.

Specifically, the main hypothesis of the project is that the capacities of the local features and the region segmentation algorithms are complementary, and thus their combination, carried out in the right way, maximizes them while minimizing their limitations. The main objective, and therefore the main contribution of the thesis, is to validate this hypothesis by proposing a framework for developing new solutions combining local features and region segmentation algorithms, obtaining solutions with improved capabilities.

As the hypothesis is proposing to combine two techniques, the validation process has been carried out in two steps. First, the use case of region segmentation algorithms enhancing local features. In order to verify the viability and success of this combination, a specific proposal, SP-SIFT, was been developed. This proposal was validated both experimentally and in a real application scenario, specifically as the main technique of object tracking algorithms.

Second, the use case of enhancing region segmentation algorithm with local features. In order to verify the viability and success of this combination, a specific proposal, LF-SLIC, was developed. The proposal was validated both experimentally and in a real application scenario, specifically as the main technique of a pigmented skin lesions segmentation algorithm.

The conceptual results proved that the techniques improve at the capabilities level. The application results proved that these improvements allow the use of this techniques in applications where they were previously unsuccessful. Thus, the hypothesis can be considered validated, and

therefore the definition of a framework for the development of new techniques with improved capabilities can be considered successful.

In conclusion, the main contribution of the thesis is the framework for the combination of techniques, embodied in the two specific proposals: enhanced local features with region segmentation algorithms, and region segmentation algorithms enhanced with local features; and in the success achieved in their applications.

Resumen

Muchas y muy diferentes son las propuestas que se han desarrollado en el área de la visión artificial para la extracción de información de las imágenes y su posterior uso. Entra las más destacadas se encuentran las conocidas como características locales, del inglés *local features*, que detectan puntos o áreas de la imagen con ciertas características de interés, y las describen usando información de su entorno (local). También destacan las regiones en este área, y en especial este trabajo se ha centrado en los segmentadores en regiones, cuyo objetivo es agrupar la información de la imagen atendiendo a diversos criterios.

Pese al enorme potencial de estas técnicas, y su probado éxito en diversas aplicaciones, su definición lleva implícita una serie de limitaciones funcionales que les han impedido exportar sus capacidades a otras áreas de aplicación. Se pretende impulsar el uso de estas herramientas en dichas aplicaciones, y por tanto mejorar los resultados del estado del arte, mediante la propuesta de un marco de desarrollo de nuevas soluciones.

En concreto, la hipótesis principal del proyecto es que las capacidades de las características locales y los segmentadores en regiones son complementarias, y que su combinación, realizada de la forma adecuada, las maximiza a la vez que minimiza sus limitaciones. El principal objetivo, y por tanto la principal contribución del proyecto, es validar dicha hipótesis mediante la propuesta de un marco de desarrollo de nuevas soluciones combinando características locales y segmentadores para técnicas con capacidades mejoradas.

Al tratarse de un marco de combinación de dos técnicas, el proceso de validación se ha llevado a cabo en dos pasos. En primer lugar se ha planteado el caso del uso de segmentadores en regiones para mejorar las características locales. Para verificar la viabilidad y el éxito de esta combinación se ha desarrollado una propuesta específica, SP-SIFT, que se ha validado tanto a nivel experimental como a nivel de aplicación real, en concreto como técnica principal de algoritmos de seguimiento de objetos.

En segundo lugar, se ha planteado el caso de uso de características locales para mejorar los segmentadores en regiones. Para verificar la viabilidad y el éxito de esta combinación se ha desarrollado una propuesta específica, LF-SLIC, que se ha validado tanto a nivel experimental como a nivel de aplicación real, en concreto como técnica principal de un algoritmo de segmentación de lesiones pigmentadas de la piel.

Los resultados conceptuales han probado que las técnicas mejoran a nivel de capacidades. Los resultados aplicados han probado que estas mejoras permiten el uso de esta técnicas en aplicaciones donde antes no tenían éxito. Con ello, se ha considerado la hipótesis validada, y por tanto exitosa la definición de un marco para el desarrollo de nuevas técnicas específicas con capacidades mejoradas.

En conclusión, la principal aportación de la tesis es el marco de combinación de técnicas, plasmada en sus dos propuestas específicas: características locales mejoradas con segmentadores y segmentadores mejorados con características locales, y en el éxito conseguido en sus aplicaciones.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

1.1 Background and motivation

Artificial Intelligence (AI) is having a major impact in today's world in varied and diverse scenarios. One of the main reasons for such impact is the exponential growth of available data, especially for image and video processing. The other is the improvement of the technologies in terms of performance and capabilities.

Ten years ago, there were just two main areas where video was broadly used: entertainment and security. Technology was differently applied on each of them. The main objective of the former was to improve video quality and reduce band used, i.e. compression. Whereas the latter was focused on automating surveillance tasks. e.g. abandoned object detection.

Despite the possibility of defining computer vision in many ways, we here understand computer vision closer to the second application. Ten years later, communications, storage capacities and capturing sensors, have improved so much that the number of areas where image is used is considerably larger. Therefore, also computer vision application fields have grown. Thus, we define:

Computer Vision is a field of Artificial Intelligence that aims at giving computers an understanding of the world through visual information.

Following the computer vision definition, the researchers' goal for building applications is to find the best strategy for teaching computers.The Figure 1.1 shows a commonly admitted organization of artificial intelligence strategies: traditional or fully hand-crafted, machine learning and deep learning [LeCun et al., 2015; Ben-David and Frank, 2009].

All three strategies share both the feature extraction and the classification stages. Feature extraction can be defined as the obtention of image elments adequated for a computer to perform

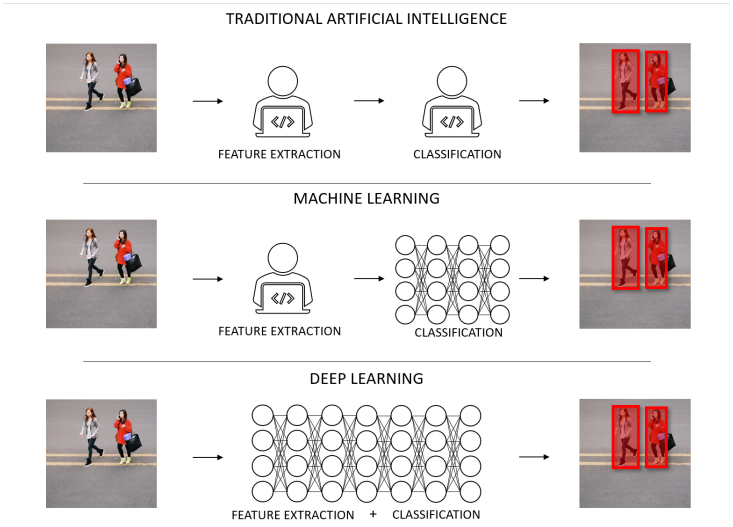


Fig. 1.1. Artificial intelligence strategies. Traditional AI uses both human-defined feature extraction and classification methods. Machine learning approaches combine human-based feature extraction with learned classification solutions. Deep learning strategies perform both feature extraction and classification as a learning processes.

a specific task, classification is usually referred as the process of associating the presence of such features with one of the expected outputs (classes). As in all serial processes, first stages are critical for success. In fact, classifiers have become more and more complex along time to solve feature extraction issues.

Feature extraction cannot be considered a novel task, but neither can be considered solved [Lowe, 1999; Ono et al., 2018]. However, its impact in the classification process is undeniable: most of the vision challenges that are considered solved or handled are rooted on successful feature extraction stages [Keel et al., 2019].

In recent years, deep learning has become really popular as a result of its success in computer vision challenges such as image classification [Russakovsky et al., 2015]. One of the keys for this success is the excellent performance of the features extracted by deep neural networks. There are several references in the literature [Zhou et al., 2014; Liu et al., 2015], where deep learning features fed into traditional classifiers perform as good as complete deep learning approaches.

There are several tasks where the available amount of data is not enough to train deep learning architectures from scratch. For these tasks, transfer learning strategies are the preferred choice [Shin et al., 2016]. However, it is not clear that features proven successful for specific tasks such as image classification, are also valid for tasks such as image registration or identification.

Machine learning will be the solution for these tasks. Notwithstanding, the impact of the feature extraction stage is critical in these approaches.

As aforementioned, feature extraction is not exactly a novel task. First relevant approaches

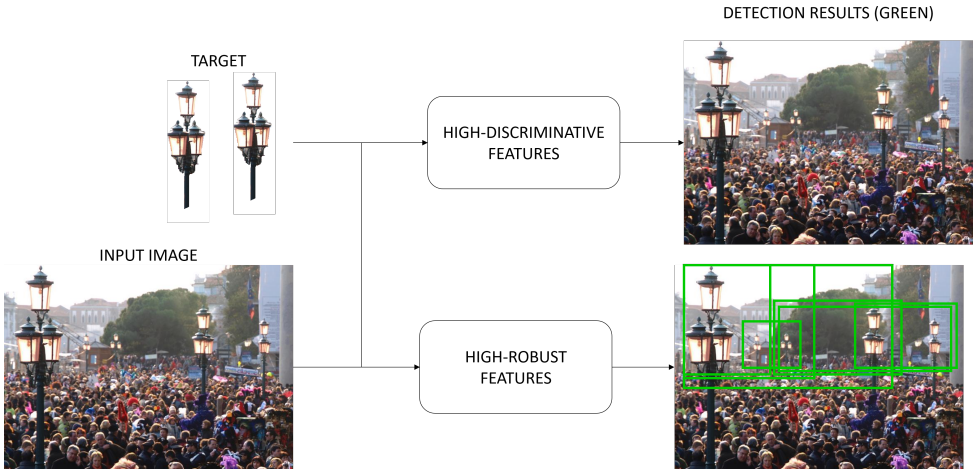


Fig. 1.2. Example of feature extraction challenges. Target objects are searched in the input image. High discriminative features have zero detections (top row). High robust features have multiple detections.

[Lowe, 1999; Shi and Tomasi, 1994] are now 20 years old. Contrasted revisions have been proposed [Tuytelaars et al., 2008], and evaluation benchmarks covering a wide variety of features are publicly available for research [Mikolajczyk et al., 2005; Mikolajczyk and Schmid, 2005].

The lack of overall success—it is unfair to say that they have not been successful in any task—may be associated to its task-oriented design. Some references consider that traditional feature extraction techniques are biased by the information used, e.g. local gradient's information [Bay et al., 2008; Tola et al., 2009] or region's color information [Manjunath et al., 2001]. However, to date, there is not a common and generic feature extraction approach that can be considered successful.

On the one hand, the variability of the information in computer vision challenges is huge. By definition, discrimination power and robustness against variability are opposite characteristics. The most robust features proposed to manage variability tend to have a very low discrimination power. On the other hand, the amount of information contained in images and video sequences increases the probability of wrongly associating information. The most discriminative features struggle when contextual information interferes with the target being described. Figure 1.2 shows and example. High discriminative features are unable to associate the target object with any element of the input image as the background information interferes. The high robust features yields several detections as they do not consider sizes nor points of view.

In summary, several feature extraction strategies appear in the literature with different levels of success. Their main limitations are design-related, but in some occasions those limitations can be identified and compensated.

1.2 Objectives

The main objective of this thesis is to explore ways of combining features extraction techniques towards more robust and discriminative solutions. In case of success, we will end up proposing new versatile schemas to build these solutions. We expect to improve existing state-of-the-art solutions in several challenges, widening their application fields.

For achieving this objective, we propose to work on the following areas:

- Feature extraction techniques. There is a wide range of feature extraction techniques in the literature. We will study the more successful solutions, analyzed their strengths and weaknesses, and define potential combination scenarios.

- Development of feature combination scenarios. The fusion of features in a post-processing fashion is a proven unsuccessful strategy. The goal is to maintain their original strengths and reduce their weaknesses. We propose to merge features at the definition level, combining their extraction techniques to maximize their potential.

- Feature combination application. We will study new application fields for these features. We will look for areas where the proposed combination's capabilities can boost the state-of-the-art performance.

1.3 Outline and major contributions

The main contributions of this thesis are summarized below:

1. We present a hypothesis for the development of computer vision tools based on the combination of local features and region segmentation algorithms.

2. We present a combination of local features and region segmentation algorithms that improves the discriminative capacity of the features and widens their fields of application [Navarro et al., 2014b].

3. We present a combination of region segmentation algorithms and local features that improves the scale-space supporting capacity of the region segmentation algorithms and widens their field of application [Navarro et al., 2018a].

4. We present two applications of the local features and regions combination, to validate the proposal and the thesis hypothesis [Navarro et al., 2014a, 2018b].

5. We present an application of region and local features combination, to validate the proposal and the thesis hypothesis[Navarro et al., 2018a].

In other level of relevance, we worked in several state-of-the-art reviews in order to propose well-judged hypothesis. The main ones are the following:

1. We analyze two wide areas of computer vision solutions, local features and region segmentation techniques [Navarro, 2014a].
2. We carrry out an exhaustive state of the art review and comparative evaluation of local features [Martín Redondo, 2016].
3. We perform a viability study of applications of our proposals [Navarro, 2014b].

1.4 Structure of the document

This document is structured in five parts, which are organized as follows:

- Part I: Introduction
  - *Chapter 1: Introduction.* This chapter presents the motivation, the objectives, the main contributions and the structure of this thesis.
- Part II: Main hypothesis: state-of-the-art review and theoretical discussion
  - *Chapter 2: Definitions and literature review.* This chapter reviews the main state of the art contributions in the two main areas studied in the thesis: local features and region segmentation algorithms.
  - *Chapter 3: Main hypothesis.* This chapter presents the thesis main hypothesis, which is divided in two statements that will be further validated.
- Part III: Enhancing the discriminative capacity of local features via region segmentation
  - *Chapter 4: The SP-SIFT local feature: Local features discriminative capacity enhanced by region segmentation algorithm.* This chapter presents the proposed schema for the combination of local features and region segmentation to enhance the discriminative capacity of the former. The schema is materialized in the proposal of the SP-SIFT local feature.
  - *Chapter 5: SP-SIFT validation via tracking applications.* This chapter presents two applications of SP-SIFT to validate the success of the proposed schema, and partially validate the thesis hypothesis.
- Part IV: Enhancing the scale adaptation of region segmentation algorithms via local features

  - *Chapter 6: The LF-SLIC algorithm: Enhancing the the discriminative capacity of a region segmentation via local features.* This chapter presents the proposed schema for the combination of region segmentation algorithms and local features to enhance the scale-space support of the former. The schema is materialized in the LF-SLIC region segmentation algorithm.
  - *Chapter 7: LF-SLIC validation via skin lesion segmentation application.* This chapter presents an application of the LF-SLIC region segmentation algorithm to validate the success of the proposed schema, and partially validate the thesis hypothesis.

- Part V: Conclusions
  - *Chapter 8: Achievements, conclusions and future work.* It concludes this document summarizing the main results and discussing potential future work for its extension.

- Part VI: Appendixes
  - *Appendix A: Publications.*
  - *Appendix B: Spanish translation of achievements, conclusions and future work.*

- Glossary
- Bibliography

The relationships between chapters and parts of the thesis are depicted in Fig. 1.3.

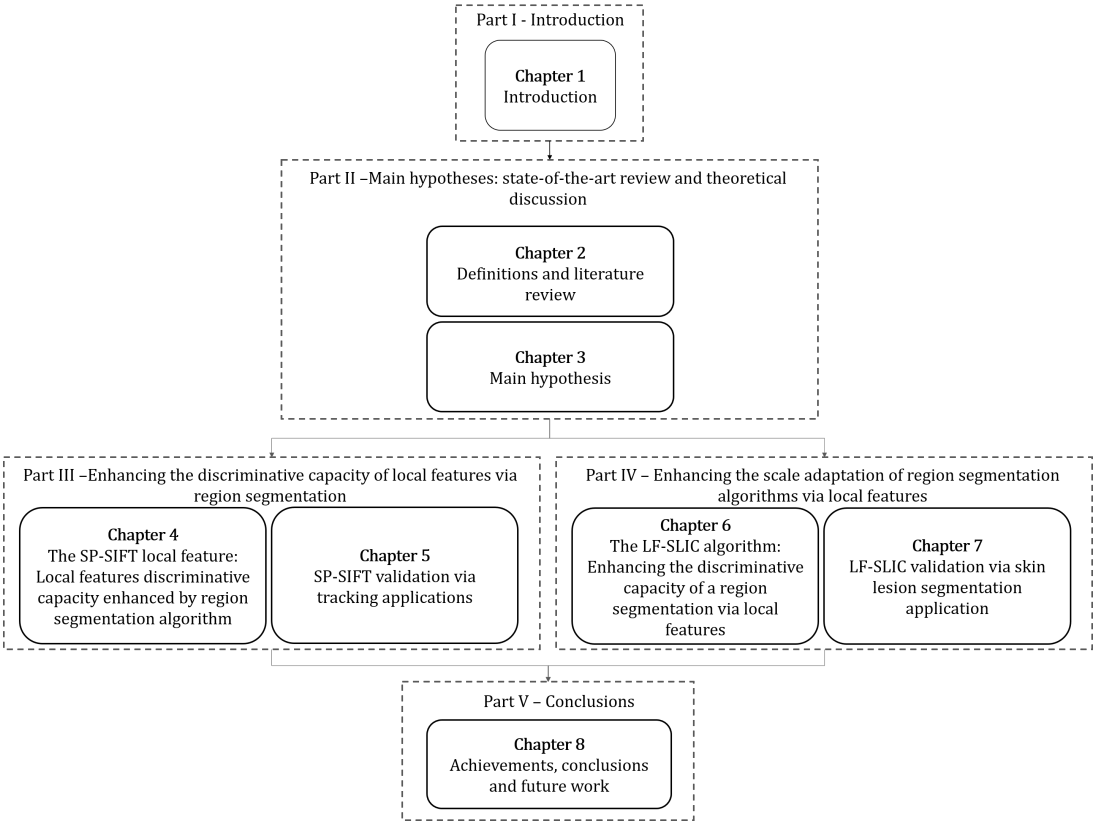


Fig. 1.3. Dependence between the chapters of this thesis.

Part II

Main hypothesis: state-of-the-art review and theoretical discussion

Chapter 2

Definitions and literature review

In this chapter we present a study of the reference methods in the literature of feature extraction for image processing. We divide existing proposals in three categories: global features, local features and region segmentation algorithms. Attending to their impact in the state-of-the-art, we discuss the main techniques in the last two categories. The objective of the chapter is to identify the key techniques on which build the thesis main hypothesis. We present the selected methods in the last section of the chapter. This chapter is built on the work developed as a scientific visitor in Queen Mary University of London in February 2014: "Local features and superpixels techniques" under the direction of Prof. Andrea Cavallaro, and on the results of a Master Thesis supervised during the PhD: "Comparative evaluation of detection and description keypoint techniques"[Martín Redondo, 2016].

2.1 Introduction to image features

This section is devoted to establishing common definitions on the concepts used throughout the document.

First, we define the concept of feature according to its source, its properties and its extraction strategy. In this document, a feature is defined as an entity obtained from the image content, with certain properties acquired by the extraction process, e.g. invariance, robustness or distinctiveness. Features are intended to be used to link semantic information to the feature spatial location in the image. The information can be for instance a relation with a similar area of another image (image stitching), a relation with a model with a semantic definition (object detection), or the relation of the whole image information to certain category (image retrieval).

Features, as we define them, can be categorized in different ways. Our definition of feature levels is based on the amount of image information used in the feature extraction. The resulting categories are then: global features, region segmentation algorithms and local features.

Global features: A global feature uses all the information in the image for description. Different global features have been proposed in the state-of-the-art like color histograms, color variations or image principal component analysis. They demonstrate reliable performances in image retrieval or object recognition with clean scenarios, i.e. with low levels of clutter, no occlusions or no background-foreground differentiation required. The range of applications and scenarios where these features are useful is very limited.

Region segmentation algorithms: A region segmentation algorithm divides and encapsulates the image information in a number of non-overlapping segments. Each segment or region is described using just the information contained in its spatial extent. This approach overcomes many of the limitations of the global features. It aims to generate different descriptions or features for the different elements in the image. Their definition can be done using any source of groupable information, e.g. color intensities, gradients, frequency, ... Unlike global features, region segmentation algorithms are used in a wide variety of applications.

Local features: A local feature can be an image pattern which somehow differs from its neighborhood. It is associated with a change of image properties. Those features can be pixel values, regions or more complex combination of properties as gradient's orientations or filter's responses. The difference between local features and the two above categories is the detection process. It only uses local information around the detection area. The number of applications for these techniques, as for the region segmentation algorithms, is very high.

The objective of the thesis is to propose new features to widen computer vision application fields, or at least improve their results. This will be achieved by obtaining features with new properties that will made them available to operate in a variety of applications. The use of global features narrows the scope; therefore, we discard them from the pool of solutions to analyze and focus only on local features and region segmentation algorithms, their respective state-of-the-art proposals, and their applications and limitations.

2.2 Local features: definition and state-of-the-art

Local features (LF) have shown to be a very useful tool in many computer vision applications with more than 2 million studies proposed during the past two decades.

According to our definition of LF, they are entities extracted, i.e. detected and described, by applying several operations on the image information. Thus, the extraction process of a local

feature is a two steps process: the detection (LF-DT) and the description (LF-DS). At each of these stages, the LF acquire different properties, that are reviewed hereinafter.

Those properties will allow to find a LF wherever it appears. i.e. in different images. The process of associating LF from different images is defined as matching.

2.2.1 Local features: detection and description

2.2.1.1 Local features detection - LF-DT

The detection stage consists of locating areas of the image that differ from its neighborhood. Different image properties can be considered in the LF-DT, e.g. colors or textures. A prototypical LF-DT process may provide several properties to the LF. However, some of them may inter-conflict and compromises need to be made. The most relevant are here described.

Repeatability is the most relevant property for a LF-DT. It encompasses a set of other properties later analyzed. It is defined as the property of a feature to be detected under different circumstances, i.e. image variations, environment changes or partial occlusions. However, increasing repeatability may lead to a lower discriminative power.

Distinctiveness is a property related both to the LF-DT and LF-DS. It is defined as the capacity of a LF to be correctly matched even after suffering strong changes as spatial, illumination or point-of-view transformations. A high distinctive LF will hardly be wrongly matched to a similar, but different feature. Low distinctive LF can be easily mismatched to similar LF after moderate changes.

Accuracy is a key property depending on the application. The capacity of a LF-DT to accurately locate several appearances of the same LF under different conditions is key for applications like image registration.

Quantity , as accuracy, is also a key property for some applications requiring a large number of LF to operate. On one hand, the use of a high number of features requires high distinctive features to avoid mismatches. Yet, this comes at the cost of increasing the computational cost required for description and posterior comparison. On the other hand, applications like camera calibration or scene understanding require a high number of features to obtain a robust calibration or compensate for missing scene information.

As aforementioned, some properties are conflicting. LF-DT with high repeatability figures are prone to avoid high numbers of LF detections which may lead to the creation of low distinctive LF. The spatial extent of a LF is not mentioned directly but can be considered a relevant property. Increasing the spatial extent of a LF results in a very distinctive feature. However, it also may lead to a low repeatable feature. We do not include it as we consider that it is a design

criterium and not a real LF property. The accuracy is not directly related to any of the other properties, but both repeatable and distinctive LF must present higher invariance to changes, and hence, provide a higher accuracy under a wider range of global and local changes. The quantity property could be directly related to repeatability and opposed to distinctiveness: the higher the quantity the more challenging the distinctiveness, and the easier the repeatability.

2.2.1.2 Local features description - LF-DS

The description step consists in the characterization of the LF. Keeping in mind the objective of the LF, the characterization intends to provide several properties to it. The final application usually addresses which LF-DS is the most appropriate to use, but also the LF-DT affects the LF-DS choice. However, it is possible to summarize the main properties desired in a LF-DS. To better establish the properties provided by the LF-DS, we will assume that LF-DS are defined together with an associated metric for matching.

Repeatability is one of the key properties for a LF-DS. First, let's assume that the same LF is detected under different conditions. The LF-DS goal is to characterize the detections in a way that the final descriptor is the same disregarding the different conditions, e.g. in two images capturing the same scene from different points-of-view. To accomplish this objective, two complementary strategies must be followed. First, it is possible to define invariant description processes. A description is invariant if the resulting descriptor is unaffected by image transformations. This can be achieved by mathematically modeling those transformations, e.g. rotations or scale changes. The second strategy is to remain unaffected by image transformation using low sensitive descriptions. For example, if the impact of noise and blur is excluded from the description, the resulting description may be similar disregarding its presence.

Distinctiveness is a property that plays a major role for many applications. The description directly depends on the nature of the LF-DT. However, it is possible to describe a high distinctive LF in a low distinctive way and vice versa. The objective here is to describe a LF differently than any other feature.

Efficiency is an application dependent property. The description process can be computationally costly, but also the detection. The difference here is the post-processing strategies that may follow. The characterization nature, e.g. the dimensions of the description vector, their binary/not binary nature and the associated metrics used to match them, severely affect the computational costs. Differences may be of a significative order, e.g. there is a substantial difference (measured in powers of ten) between matching 128-dimensional and non-binary vectors using the Euclidean distance and matching 64-dimensional binary vec-

tors using the Hamming distance. Therefore, real-time or high-quantity-LF applications require efficient LF-DS.

On the basis of these properties, several conclusions arise. Repeatability includes invariance and robustness strategies. Enhancing one of them affects directly to the discriminative capacity of the description. The higher the robustness the lower the distinctiveness and vice versa. Moreover, the higher the invariance the higher the accuracy required by the detection process. The invariance is also highly linked with the efficiency. Most of the times, the higher the complexity of the image transformations modeled in the description process the higher invariance but also the lower the efficiency.

2.2.2 State-of-the-art in the local features

The LF field is one of the most extensively studied ones in computer vision. Despite the broad number of LF techniques reported in the literature, their automatic characterization and matching is still an unresolved issue. We present now a summary of the evolution of the techniques, a brief description of the key references of the state-of-the-art, and a property-based comparison.

To talk about point features is to talk about one of the most referenced papers in computer vision: SIFT [Lowe, 1999]. The SIFT feature became a standard for LF, both in techniques, and in performance. However, it presents several drawbacks, some of them related to computational costs.

In the last two decades, a variety of solutions were proposed, e.g. [Bay et al., 2006; Calonder et al., 2010; Leutenegger et al., 2011; Rublee et al., 2011]. Despite most of them reported improvements in some properties: invariances, accuracy or efficiency, no method has replaced SIFT as the reference technique. Most of these techniques present solutions to very specific SIFT or LF issues, but rarely none faces from scratch the overall detection and characterization problem.

However, analyzing the state-of-the-art, it is possible to define three main development lines that have driven the evolution of LF. In order to perform a state-of-the-art discussion, we first define what we consider the three major breakthroughs in the LF development and situate them in time.

1. SIFT appearance. When SIFT [Lowe, 1999] and its following improvement [Lowe, 2004] appeared, its results edged above all the previous techniques. This can be considered as starting point of the LF widespread use. Its novel approach for information extraction changed the paradigm. Its main contribution may be the scale-invariant use of histograms of oriented gradients (HOG) for the features description. However, the combination of all the techniques: difference-of-Gaussians (DoG) for the detection, histograms of gradients for the description and the matching metric, is what really set the standard. Before SIFT,

there had been "singular points" detectors proposed in the literature. We define singular points as those pixels or areas of the image that stands out in their environment attending to a specific property, e.g. color. The best known pre-SIFT detector is the Harris detector [Harris et al., 1988]. Using Harris detector, some local feature techniques were proposed before SIFT. The most relevant, because served as an inspiration for SIFT are [Zhang et al., 1995] and [Schmid and Mohr, 1997]. The first one applied the Harris detector to optimize epipolar correlation by comparing only the points where a Harris feature is detected. The second one proposed a SIFT like description for each Harris point: an orientation-invariant vector of derivative-of-Gaussian image measurements. Both were lately outperformed by SIFT. The first one in the description technique, the second one by including scale aware methods in the proposal.

2. Fast features. If it were not for the SURF LF [Bay et al., 2006] this stage would be called fast binary descriptors. SIFT results in terms of accuracy and repeatability were good enough to widen the LF application fields. As a consequence, new requirements appear in terms of computational and temporal costs. To solve these issues, a new branch of developments appeared. Initially leaded by the SURF feature, a bunch of techniques were developed to speed-up SIFT-like LF. The SURF feature proposed the use of integral images [Viola et al., 2001] to speed-up the detection stage, proposing the "Fast-Hessian" and Haar-wavelet responses again with integral images to speed-up the description stage. It achieved sensible reduction in time costs at a very low cost in accuracy and robustness. However, faster techniques were required soon. Binary descriptors [Alahi et al., 2012; Calonder et al., 2010; Leutenegger et al., 2011; Rublee et al., 2011] aimed to fill this gap. The idea behind binary descriptors is reducing the computational cost associated to the matching process. They propose descriptors where each bit is independent and the much faster Hamming distance can be used as similarity measure instead of more expensive distances, e.g. the Euclidean distance.

3. Learning approaches. There have been several unsuccessful approaches to learning LF, but only with the recent deep learning boost they start working as expected. Just like previous improvements, biggest efforts were focused on descriptors. ConvOpt [Simonyan et al., 2014], presented in 2014, proposed to learn the spatial pooling regions of the descriptor. In DeepDesc [Kumar et al., 2016] and TFeat [Balntas et al., 2016], they proposed deep learning architectures to extract gradients. Similar to SIFT, LIFT is defined as learned invariant feature transform [Yi et al., 2016]. It presents an end-to-end feature, i.e. detector and descriptor. It is based on a combination of CNNs (Convolutional Neural Networks). A well-known example of a learned detector is TILDE [Verdie et al., 2015].

State-of-the-art main proposals

We follow a chronological order in the algorithm analysis. We also describe here those algorithms that will be used for the evaluations in the further chapters of this thesis. For selecting the algorithms, we followed two criteria: contrasted references and evaluations in the state-of-the-art, and variety in the proposed main contribution of each technique. Key jobs for the thesis development will be extended. Notice that among the following techniques there are LF-DT, LF-DS and full methods.

Hessian corner [Beaudet, 1978]

- Type: LF-DT.
- Key contributions: It performs a search using the spatial derivatives, specifically the second derivatives. The resulting detections are those locations in the image with strong second derivatives in two orthogonal directions.
- Discussion: The proposal obtains the best results in situations where scale doesn't change. A scale invariant version has been proposed: Hessian-Laplace [Mikolajczyk and Schmid, 2001]. Also an affine invariant version has become famous: Hessian-affine [Mikolajczyk and Schmid, 2002]. Those last two provide blob-like detections, with a remarkable capability of avoiding edges as compared to following techniques as difference-of-Gaussians (DoG).

Harris corner [Harris et al., 1988]

- Type: LF-DT.
- Key contributions: The proposal is an extension of Moravec's corner detector. It performs a spatial patch level analysis of image pixels. It extracts the intensity of the pixels in a patch and measures the variation of shifting the patch around. It the sum of the intensities in the patch increases in several shift directions, it indicates the central pixel is a corner.
- Discussion: The proposal obtains the best results in tracking and stereo matching applications, i.e. situations where scale doesn't change. A scale invariant version has been proposed: Harris-Laplace [Mikolajczyk and Schmid, 2001]. Also an affine invariant version has become famous: Harris-affine [Mikolajczyk et al., 2005].

SIFT [Lowe, 2004]

- Type: Full method.

- Key contributions: First, it proposes a detector. The detector applies the DoG on the scale-space extrema to find detection candidates. Then, it defines a number of stability criteria to filter unstable detections, e.g. detections in weak edges. Second it proposes a gradient distribution descriptor. In a first stage, it obtains the features main orientation. In the second stage, it creates the LF-DT using the gradient magnitude and orientation in the region around the detection. It also proposes a feature matching criterion based on relative Euclidean distance between pairs of descriptors.

- Discussion: As a result of its detection process, points are detected repetitively under changes of illumination, blurring, scale and rotation. The proposed descriptor is robust against most of the image changes as illumination, rotation and scale. It was defined as invariant to viewpoint changes and affine transformations of the image, but its capacity is constrained to limited values. Apart from the limitation on viewpoint changes, the description using the environment information could lead to wrong description under occlusion or variable backgrounds. The computational cost is the last and probably the most tackled weakness of the method for some applications.

GLOH [Mikolajczyk and Schmid, 2005]

- Type: LF-DS.

- Key contributions: It proposes an extension of the SIFT descriptor by changing the grid for a log-polar location one. It also proposes to use Principal Component Analysis (PCA) to decrease the size of the descriptors, thus increasing its robustness and distinctiveness.

- Discussion: The method is invariant to the same changes than SIFT. The modification of the grid leads to more stable descriptions. Blurring is the only invariance in which the GLOH performance is expected to be below SIFT's performance.

SURF [Bay et al., 2006]

- Type: Full method.

- Key contributions: The detection stage of SIFT is changed in SURF. It uses integral images to efficiently compute a rough approximation of the Hessian matrix. The process is computed per different scales. The speed up in the description process is obtained by describing the features with the response of a few Haar-like filter.

- Discussion: Gaussians are optimal for scale-space analysis, but in practice they have to be discretized introducing artifacts, in particular for small Gaussian Kernels. However, in spite of the rough approximations, the performance of the feature detector is comparable

to the results obtained with the discretized Gaussians. In the description, SURF does not apply SIFT's spatial weighting scheme, which produces damaging artifacts.

DAISY [**Tola et al., 2009**]

- Type: LF-DS.
- Key contributions: It proposes a descriptor improvement by switching SIFT's weighted sums of gradient norms by orientation maps that are convolutions of the gradients with several directional Gaussian filters.
- Discussion: DAISY appears to face image matching on cluttered environments. Despite the results in crowdy environments, features detected in non-crowdy situations are described with less discriminative descriptors. In summary, DAISY only overcomes SIFT in some specific scenarios, but alleviates the computational cost.

BRIEF [**Calonder et al., 2010**]

- Type: LF-DS.
- Key contributions: It is one of the pioneers in binary description. It extracts binary strings from patches of interest regions instead of extracting gradient-based feature vectors. Specifically, BRIEF proposes to use binary values based on a set of intensity tests, i.e. each bit is computed by comparing the intensity difference between a pair of sample points from the image patches. Matching between the resulting binary descriptors is performed much faster than between non-binary descriptors.
- Discussion: Despite the clear advantage in computation and storage, it still has weakness in terms of reliability and robustness. A number of interesting proposals appear based on BRIEF. ORB [Rublee et al., 2011] proposes an orientation assignment and a learning method to optimize the positions, improving the results, mainly in orientation. BRISK [Leutenegger et al., 2011] is similar to ORB, but proposes to use a DAISY-like sampling pattern to improve performance in cluttered environments.

KAZE [**Alcantarilla et al., 2012**]

- Type: LF-DT.
- Key contributions: It proposes to improve the scale support provided by Gaussian approaches like SIFT's. Instead using the Gaussian scale, which does not respect the natural boundaries of objects and smooths to the same degree both details and noise, it proposes a nonlinear scale space support by means of nonlinear diffusion filtering.

- Discussion: The proposal obtains a blurring locally adaptive to the image data, reducing noise but retaining object boundaries, obtaining superior localization accuracy and distinctive capacity. However, its computational cost is superior to previous approaches like SURF.

FREAK [Alahi et al., 2012]

- Type: LF-DS.
- Key contributions: The FREAK sampling pattern mimics the retinal ganglion cells distribution with their corresponding receptive fields, resulting in a very similar sampling pattern to the DAISY. Similar to BRISK, the LF orientation is added from local gradients, but FREAK uses pairs with symmetric fields around the receptive center. The descriptor is constructed by thresholding differences of corresponding Gaussian kernels.
- Discussion: This technique was a breakthrough due to a remarkable trade-off between computational and memory cost, and its ability to provide high invariant LF-DS. However, worse results obtained in datasets different from the one used in the design have lowered the expectations.

TILDE [Verdie et al., 2015]

- Type: LF-DT.
- Key contributions: It proposes a learned detector. It defines positive and negative detections, using reference detectors as SIFT or SURF. Using the proposed classification, it trains a regressor that provides a value for different image patches. Negative patches obtain lower values whereas positive ones obtain higher values. Local maxima are applied to obtain final detections.
- Discussion: Results overcome previous approaches at the cost of increasing figures in false positive ratios. However, its major drawback is related to training data. The technique achieves great performance when tested in data similar to the training. Its results are worse in unseen conditions. This is a consequence of the lack of generality of learned LF.

DeepDesc [Kumar et al., 2016]

- Type: LF-DS.
- Key contributions: It proposes to learn the filter banks to extract gradients with fixed pooling. It uses a Siamese architecture, i.e. two 3-layer CNNs that share weights, guided

by the minimization of a Contrastive loss. The Contrastive loss is used to maximize the difference between positive and negative samples.

- Discussion: Results overcame previous approaches in most common benchmarks. However, high computational costs and the requirement of a training stages are its most relevant issues.

TFeat [Balntas et al., 2016]

- Type: LF-DS.
- Key contributions: Similar to DeepDesc, it proposes to learn the filter banks. The key difference is the learning architecture. It proposes to use a triplet network, where each net is a CNN with 2 convolutional layers and a fully connected layer. The learning is guided by the minimization of either the Margin ranking loss or the Ratio loss, both searching for a simultaneous minimization of the distance between positives samples and a maximization of the distance between negative samples.
- Discussion: The computational costs are about the half of other learned features as DeepDesc. Its performance equals DeepDesc in learned scenarios but lacks generality in unseen environments.

LIFT [Yi et al., 2016]

- Type: LF-DS.
- Key contributions: It defines three components to be trained: detector, orientation estimator and descriptor. First, it trains a Siamese network based on TILDE for the detection process. Then, the orientation predictor is trained using a Siamese network guided by the Euclidean distance between rotated versions of the description patches. Finally, for the descriptor, it trains a DeepDesc-like architecture.
- Discussion: In overall obtains better repeatability than previous learned approaches. However, in the presence of viewpoint changes, LIFT performs poorly since it aims only for translation invariance and not for scale or affine invariance.

2.3 Region segmentation algorithms: definition and state-of-the-art

Region segmentation is the process of dividing the whole space into non-overlapping regions. As a result, no part of the space can be leaved unassigned. This technique is employed as a preprocessing step to annotation, enhancement, classification and/or information extraction.

Region segmentation has been widely discussed in the last 20 years and is still today a key technique in some of the aforementioned applications under the deep learning paradigm. As a consequence, the are several studies and surveys on the topic [Salembier and Marqués, 1999], [Freixenet et al., 2002], [Ilea and Whelan, 2011] or [Escudero-Viñolo, 2016] which inspire the following region segmentation algorithms classification.

- Local region segmentation algorithms: clustering, region merging and mode-seeking
- Global region segmentation algorithms: energy minimization problems.
- Combined approaches: graphs for global representation of local information.

Before deepening into the region segmentation algorithms definition and state-of-the-art techniques, we are going to narrow this broad area that are the region segmentation algorithms. To do so we considered three main aspects: up-to-date techniques yielding top state-of-the-art contrasted performance . However, it is important to bear in mind that this thesis hypothesis and results can be extrapolated to other LF techniques and to other region segmentation methods.

Under these restrictions, superpixels techniques are the best candidate to narrow the region segmentation algorithms state-of-the-art. They include techniques in all the three above categories, every year new superpixels approaches appear in the state-of-the-art, and their results are fairly contrasted.

2.3.1 Region segmentation algorithms—superpixels—definitions

As defined in [Achanta et al., 2012], superpixels are groups of pixels with perceptual meanings. A number of algorithms and computer vision applications have been built on top of superpixels algorithms, as they provide a significant reduction in the information to be processed compared to the pixel-level, while preserving key information of the image. According to [Wei et al., 2018], a successful superpixels method should provide the following properties:

- **Adherence to boundaries**: Superpixels should adhere well to image boundaries such that each superpixels only overlaps with one object.
- **Computational efficiency**: The computational complexity for an efficient segmentation algorithm should be independent of the number of superpixels and linear/sublinear with the image size.
- **Hierarchical segmentation**: Superpixels segmentation results at different levels should be close to the human vision system. Numerous algorithms can benefit from multi-resolution representations of images and hierarchical superpixels can be used for these tasks.

- **Preserved topology**: Superpixels should conform to a simple topology such that neighborhood relationships can be maintained.

Thanks to those capabilities, superpixels have been used in a number of applications in recent years such as medical image segmentation [Wu et al., 2014], motion segmentation [Ayvaci and Soatto, 2009], multi-class object segmentation [Fulkerson et al., 2009], target tracking [Wang et al., 2011] and object detection [Shu et al., 2013]. However, most of the state-of-the-art proposals are only focused on improving adherence to boundaries and computational efficiency, while hierarchical segmentation and preserved topology are properties usually provided by post-processing stages...

2.3.2 Region segmentation algorithms—superpixels—state-of-the-art

There is a rich literature on image segmentation. In this section, we discuss the most relevant methods to this work. First, we propose a superpixels state-of-the-art organization, and then we describe the methods on these categories.

For the organization, we propose to combine the proposals from three state-of-the-art studies: SLIC state-of-the-art comparison [Achanta et al., 2012], and the recent state-of-the-art reviews [Wang et al., 2017] and [Stutz et al., 2018]. They define two main groups: graph-based approaches and gradient-ascent—or clustering—approaches. In [Stutz et al., 2018], authors propose a more detailed classification of the approaches in the gradient-ascent category. We follow these categories and examples of each one.

1. Graph-based: They treat each pixel as a node in a graph. Similarities between neighboring pixels are defined as edge weights. Superpixels are generated by minimizing a cost function defined over the graph. Examples analyzed are N-cut [Shi and Malik, 2000], EGraph [Felzenszwalb and Huttenlocher, 2004], PB [Zhang et al., 2011] and ERS [Liu et al., 2011].

2. Gradient ascent: They group pixels into clusters, i.e., superpixels, and iteratively refine them until some convergence criteria are satisfied. Most relevant sub-categories in these group are the following:

   (a) Watershed-based - Watershed [Vincent and Soille, 1991].

   (b) Density-based - QuickShift [Vedaldi and Soatto, 2008].

   (c) Contour evolution - TurboPixels [Levinshtein et al., 2009].

   (d) Clustering-based - SLIC [Achanta et al., 2012] and LSC [Li and Chen, 2015].

   (e) Energy optimization - SEEDS [Van den Bergh et al., 2012].

State-of-the-art main proposals

We follow the previous organization as it almost follows a chronological order in the apparition of the methods. Notice that we followed a similar criterion in the method selection than that followed for the LF, i.e. we selected the methods based on contrasted references and top performances in their categories.

N-cut [**Shi and Malik, 2000**]

- Type: Graph-based.
- Key contributions: It is based in texture and contour cues to create the initial graph. The cost function is defined based on edge connections to the nodes.
- Discussion: It obtains good regularization but performs poor in terms of adherence and computational efficiency.

EGraph [**Felzenszwalb and Huttenlocher, 2004**]

- Type: Graph-based.
- Key contributions: It creates the graph based on evidences of image boundaries. Superpixels are obtained by finding the minimum spanning tree using Dijkstra's algorithm to compute the shortest paths.
- Discussion: It improves N-cut boundary adherence but resulting superpixels have irregular sizes and shapes.

PB [**Zhang et al., 2011**]

- Type: Graph-based.
- Key contributions: It reduces the problem to a binary labeling of the pixels. Each pixel can be associated to two labels. The labeling problem can be enunciated as a binary labelling problem on Markov Random Fields (MRFs). It uses the elimination function to optimize the two Pseudo-Boolean function composing the objective function.
- Discussion: The algorithm computational cost is independent of the number of superpixels, which is a great improvement. Superpixels sizes and shapes are regular.

ERS [Liu et al., 2011]

- Type: Graph-based.
- Key contributions: It uses entropy rate of a random walk on a graph as the criteria to generate superpixels.
- Discussion: The algorithm presents a top state-of-the-art boundary adherence at the cost of a poor computational efficiency.

Watershed [Vincent and Soille, 1991]

- Type: Gradient ascent-watershed.
- Key contributions: The algorithm locates image local minima. Then, it performs a gradient ascent in order to obtain watersheds, i.e. lines that separate catchment basins.
- Discussion: It presents a poor boundary adherence but a good computational efficiency. Resulting superpixels are irregular. An evolution of the algorithm, compact watershed [Neubert and Protzel, 2014], improves the method in terms of superpixels compactness.

QuickShift [Vedaldi and Soatto, 2008]

- Type: Gradient ascent-density.
- Key contributions: It is a mode-seeking technique similar to the well-known algorithm mean-shift. It uses the Parzen distance to decide the direction of the movement of the points in the feature space.
- Discussion: It ranks average in terms of boundary adherence. Despite it is much faster than the mean-shift technique, it still being costly in terms of computational efficiency. It is not possible to control the superpixels sizes and shapes.

TurboPixels [Levinshtein et al., 2009]

- Type: Gradient ascent-contour evolution.
- Key contributions: It uses the geometric flow, i.e. local image gradient, to control the superpixels growth. It stars with a lattice structure of compact regions, and dilates each seed using the geometric flow.
- Discussion: The size and compactness of the superpixels is controlled. Boundaries adherence performs poorly but can be improved by adapting the compactness and size of the superpixels.

SLIC [Achanta et al., 2012]

- Type: Gradient ascent-clustering.
- Key contributions: It initializes the superpixels centers "randomly" and then uses the assignment to redefine the location. Pixels are associated to cluster center using nearest neighbor. After the association, the cluster center is updated. The process is repeated until the error converges. Post-processing steps reassign disjoint pixels.
- Discussion: It presents a great balance between efficiency and boundary adherence, ranking top in both categories in the state-of-the-art. Superpixels are regular.

LSC [Li and Chen, 2015]

- Type: Gradient ascent-clustering.
- Key contributions: It can be presented as an evolution of the SLIC technique. It performs an iterative clustering center update and pixels association until the error converges. The difference is the information used to perform the clustering. LSC maps the image pixels to weighted points in a ten-dimensional feature space. All the operations are performed in the feature space.
- Discussion: It preserves global properties of the image better than SLIC and the superpixels tend to be more regular. The computational cost is increased.

SEEDS [Van den Bergh et al., 2012]

- Type: Gradient ascent-energy optimization.
- Key contributions: It defines an energy function in terms of color and boundaries. An initial partition of the image is performed. Then, superpixels are modified according to the energy function. The energy function is solved by a hill-climbing optimization.
- Discussion: Depending on the iteration time, SEEDS can be close to run in real-time. However, superpixels are irregular and difficult to control.

2.4 Discussion of local features and superpixels methods

The description of the different methods analyzed in this chapter included a brief discussion regarding their strengths and weaknesses. These descriptions are based on the proposals' references, and on state-of-the-art analysis. For that reason, this section presents only a comparison

between the different methods, but without any further discussion about the methods individually. As a conclusion of this section, two methods are proposed to be the base of the different analysis and proposals along the thesis. The selection does not intend to select the best method on each category nor a method with some specific properties. The objective is to select methods that works for the thesis objectives. For that reason, we look for methods with the three following characteristics:

1. Good overall results: we look for methods with proven good results in overall. We discard all the task-oriented methods, and all the proposals with excellent results in some categories at the cost of others.

2. Representative of their categories: we look for techniques that present the main aspects of their respective categories. We discard all the specific adaptations, combinations or extremely disruptive proposals.

3. Evolutions and applications: we look for methods that have been evolved or applied before, so we can compare and evaluate the solutions properly. We avoid unreferred methods or too new proposals that have limited impact.

The objective is to select the methods that are the best suited to validate the thesis hypothesis. All the discussion and the qualitative conclusions are built over the quantitative evaluations performed in [Martín Redondo, 2016; Navarro and Cavallaro, 2014] or analyzed in [Achanta et al., 2012; Wang et al., 2017; Stutz et al., 2018].

2.4.1 Discussion of local features methods

This section results are based on the work [Martín Redondo, 2016; Navarro and Cavallaro, 2014]. First we analyze the methods in terms of the properties presented in 2.2. To do so, we summarize the quantitative results presented in [Martín Redondo, 2016] included in Appendix C, and perform a qualitative analysis based on it. Second, we analyze them in terms of standard methods and evolutions presented. Finally, we select the method which best meets the presented criteria.

2.4.1.1 Techniques performance evaluation

We proposed a LF evaluation framework. We designed and recorded a database. We evaluated a number of LF-DT and LF-DS, more than those analyzed here, and extensively discussed the results. Some techniques appeared later than the evaluation. To compare them in the qualitative analysis we use state-of-the-art benchmarks.

Database

Isolated transformations	Global	Illumination change
		Viewpoint change
		Blur
	Target	Illumination change
		Blur
		Partial shadowing
Combined transformations	Global	Illumination change + Viewpoint change
		Illumination change + Blur
		Viewpoint change + Blur
		Scale + Rotation
	Target	Illumination change + blur
		Partial shadowing + blur

Table 2.1: Image transformations included in the proposed database. Global transformations affect to the whole image. Target transformations are applied only to the target whereas the rest of the image remains unaffected. A total of twelve categories are proposed.

Fig. 2.1. Overview of the proposed database.

Despite there are some reference databases in the state-of-the-art [Mikolajczyk and Schmid, 2005], we decided to propose a new database for the evaluation. Previous databases we outdated and algorithms were overfitting to them. Additionally, they evaluate methods capabilities isolated. We decided to go a step further and evaluate capabilities both isolated and combined. We also decided to use high definition images and high precision annotations.

The structure of the database is similar to the one proposed in [Mikolajczyk and Schmid, 2005]. We define a number of categories, i.e. transformations applied to the image, and for each category we provide a reference image and five samples with an increasing level of complexity in the transformation applied. Table 2.1 presents the image transformations included in the database to evaluate the LF capabilities.

Figure 2.1 presents a sample of the database. Each image correspond to one of the twelve categories presented in Table 2.1.

More details about how each transformation is defined can be found in the original work [Martín Redondo, 2016].

Metrics

Different evaluation methodologies were defined for the LF-DT and LF-DS.

LF-DT: Reference LF detections are located in the reference image. Then, for each of the five images of the category, LF detections are obtained and projected on the reference image. The following labels are defined depending on the projection results. True Positive (TP) occurs when a LF detection is projected on the same location of a LF detection of the reference image. False Negative (FN) occurs when a LF detection of the reference image receives no projection, i.e. the detection is lost when the image is transformed. Using those labels, we apply the Recall metric, $RECALL = \frac{\#TP}{\#TP+\#FN}$, to evaluate the LF-DT performance on the database.

LF-DS: Reference LF detections are located in the reference image. We project those detection in the rest of the images of the same category. Detections on all the images are described using the LF-DS. Then, descriptions of the reference image are associated, matched, with descriptions on each of the images of the same category. Depending on the results of the matching process, we define the following labels. True Positive (TP) occurs when a matching of descriptors associate the same detection on both images. False Positives (FP) are those matchings between different detections. Using those labels, we apply the 1-Precision metric, $1 - PRECISION = \frac{\#FP}{\#FP+\#TP}$, to evaluate the LF-DS performance on the database.

Results summary

In Table 2.2 we present the comparison in terms of the properties described in Section 2.2. As aforementioned, this analysis is based on the state-of-the-art study and the performance evaluation carried out. All the quantitative results supporting this analysis are included in Appendix C.

2.4.1.2 Standards and evolutions analysis

We can extract three techniques that stand-out among the methods. The first one is LIFT. It presents the best results in terms of properties. However, the technique presents two major drawbacks related to this thesis objective. First, as every learned method, is environment dependent. As one of the objectives of this thesis is to find methods that combined can widen the LF application scope, these techniques are discarded. Second, despite following the two main steps LF-DT and LF-DS, it is far from representing the traditional schema of LF.

The second technique that stands out is FREAK. Its performance in state-of-the-art benchmarks is outstanding. However, its results are not so good when used in real applications. As a

Methods	Repeatability						Distinctiveness	Locality	Quantity	Accuracy	Efficiency
	Invariance				Robustness						
	Light	Rotation	Scale	Viewpoint	Blurring	Compression					
	LF-DS	LF-DS	LF-DT+LF-DS	LF-DS	LF-DT+LF-DS	LF-DT+LF-DS	LF-DS	LF-DS	LF-DT	LF-DT+LF-DS	LF-DT+LF-DS
Hessian	-	-	+	-	++	++	-	-	+++	+++	+++
Harris	-	-	+	-	++	++	-	-	+++	++	+++
SIFT	+++	++	+++	+	+++	+++	++	++	++	+++	+
GLOH	+++	++	+++	++	++	+++	++	++	-	+++	+
SURF	+++	++	++	+	++	+++	++	++	++	++	++
DAISY	+++	++	++	+	++	+++	+++	+++	-	+++	+
BRIEF	+++	+	++	+	++	+++	++	++	-	++	+++
KAZE	-	-	+	-	+++	+++	-	-	+++	+++	+
FREAK	+++	+++	+++	++	++	+++	++	++	-	++	++
TILDE	-	-	++	-	+++	+++	-	-	++	+++	-
DeepDesc	+++	++	+++	+++	+++	+++	+++	++	-	+++	+++
TFeat	+++	++	+++	+++	+++	+++	+++	++	-	+++	+
LIFT	+++	+++	+++	+	+++	+++	+++	++	++	+++	++

Table 2.2: LF comparison. The number of symbols (+) indicates the performance of the method. (+++) means top performance, whereas (+) means lowest performance. (-) means the property does not apply.

consequence, there is a limited number of applications using FREAK as a LF-DS method.

The third and final technique is SIFT (and SURF). Both techniques present a great balance in terms of capabilities. They are considered standard techniques, and there is a huge number of applications and evolutions of both methods.

2.4.1.3 Method selection

We decided to select the SIFT method. The main reason, apart from its relevance in terms of references, is the accuracy of the technique. SURF performs a number of generalizations, that has minor impact in the results, to improve the computational costs. As the computational cost is not critical for the thesis, we decide to use SIFT.

2.4.2 Superpixels algorithms discussion

This section results are based on the work [Navarro, 2014a]. We studied there the main superpixels techniques proposed in the state-of-the-art in the light of different evaluation benchmarks. As new methods have appeared, we included recent benchmarks and evaluations in the analysis.

The evaluation criteria for superpixels are too heterogeneous. For the comparison, we use the characteristics defined in the subsection 2.3. We do not include the preserved topology as it is not an extended criterium in the state-of-the-art benchmarks, and it will be difficult to properly evaluate it. In the following lines we relate those characteristics with the metrics used to measure them.

1. Segmentation results. In this group we aim to show how is the method performance in terms of boundary adherence. Typical metrics here are precision, recall, f-score or segmentation covering.

Methods	Segmentation	Superpixels	Efficiency
N-Cut	++++	++	++
EGraph	++++	+	++
PB	++++	++	++++
ERS	+++++	+++	++
Watershed	++	++	++++
QuickShift	+++	++	++
TurboPixels	++	+++	+
SLIC	++++	+++++	+++++
LSC	+++++	++++	+++
SEEDS	+++++	+++	+++++

Table 2.3: Superpixels segmentation algorithms comparison. The number of symbols (+) indicates the performance of the method. (+++++) means top performance, whereas (+) means lowest performance.

2. Superpixels results. In this group we include all the metrics referred to how the superpixels are extracted and how much it is possible to adapt them, i.e. the hierarchical characteristic. Examples of metrics are the under-segmentation error, compactness or regularity index.

3. Efficiency results. In this group we include the metrics related to computational cost.

Different from the LF discussion section, for the superpixels discussion we did not perform an state-of-the-art performance evaluation. Superpixels state-of-the-art evaluations, as opposed to LF state-of-the-art evaluations, are recent and up-to-date. We use as the reference study [Achanta et al., 2012], and perform a comparison per property based on its results. The comparison is presented in Table 2.3.

There are three techniques that outperforms the rest of methods. SLIC, LSC, and SEEDS.

- LSC is a new method, with few references and applications in the state-of-the-art compared to SLIC or SEEDS.
- SEEDS and SLIC has equivalent performances.
- SLIC has better performance in terms of superpixels number, compactness and regularity compared to SEEDS.

We select SLIC as one the goals of the thesis is to combine techniques and modify its original implementations. Also, the number of references and the public available implementations support the selection of SLIC.

Chapter 3

Main hypothesis

In this chapter we formulate the main hypothesis of the thesis. We present the research questions arising from the state-of-the art review. As a result, we identify complementary capabilities on the analyzed techniques. We present the main hypothesis, i.e. the feature extraction strategy, and two specific schemas to validate the hypothesis.

3.1 Research question

The discussion in Chapter 2 helped to identify the solution schemas per area, LF and region segmentation, as well as the reference methods that are best suited for our purposes. Additionally, as a result of the analysis, a number of weaknesses were identified. We can group them in two categories:

1. Method-related weaknesses. We group here all the issues related with how the individual methods are defined, e.g. computational cost of SIFT descriptor [Lowe, 2004], or low accurate boundaries of the watershed color-based segmentation algorithm [Shafarenko et al., 1997] at the camouflaged-areas.

2. Strategy-related weaknesses. We refer to the issues intrinsic to the strategy, i.e. intrinsic to how the LF are defined, or how the region segmentation algorithms are defined.

The objective of the thesis is to propose combination schemas to mitigate the strategy-related weaknesses. We enunciate first the major strategy-related weaknesses identified in Chapter 2. Then the research questions are exposed. We aim to answer those research questions with the thesis hypothesis.

3.1.1 Strategy related weaknesses

Local features: They are defined as local, so its discriminative capacity is an intrinsic weakness. To compensate this weakness, they all include some kind of description of their surrounding area. The information used in that description is the most critical aspect of the LF. Too much information leads to non-repetitive LF. Too less leads to non-discriminative LF. Even the format of the description is critical for the resulting LF. There is a huge number of different proposals in the state of the art of LF, but none of them is able to handle the discriminative-repetitive tug-off-war.

Region segmentation algorithms: Depending on how the regions are obtained, a region segmentation method may generate: big and high discriminative regions at the cost of boundary segmentation quality, or smaller and less discriminative regions highly tighten to the real object boundaries. Smaller ones are the preferred option in recent times. However, this is because they add additional cues on top of the region segmentation, but not because they solve the problem.

3.1.2 Statement of research questions

In a progressive approach, we enunciate the researching questions per area, to end up coming together in a main thesis research question.

At local feature level the question would be: is it possible to, somehow, constrain the information to be used in the LF description? If so, we will be able to include information in the description process to increase its discriminative capacity but at the same time avoiding the inclusion of non-repetitive information, i.e. variable or low reliable information. At a higher level, the question can be simplified in: *is there a solution to improve both discriminative and repetitive capacity of the LF at the same time?*

At region segmentation level the question would be: is it possible to define a mechanism to automatically choose whether the regions should grow or stabilize? If so, we will be able to enlarge and increase discriminativeness in homogeneous or non-relevant areas of the image, and also to stay small and adapt to highly detailed boundaries (discriminative *per se*). At a higher level, the question can be simplified in: *is there a solution to improve regions discriminative capacity and improve boundary segmentation results?*

We put together the specific research questions, and the different technologies capabilities, to expose the main research question of the thesis:

Are the main capabilities of LF and region segmentation algorithms complementary?

3.2 Main hypothesis statements

The project main hypothesis is the proposed answer to the main research question exposed in the previous section. Its wording is the following:

The combination of LF—based on local descriptions—and region segmentation algorithms—based on spatial and color relationships—results in a new family of image features with a wider application field thanks to their complementary strengths: discriminative capacity and robustness to variations in the local information.

The hypothesis opens here two possible combinations that must be validated to consider the hypothesis proven. The two combinations are the following:

1. Define a combination schema to *improve both discriminative and repetitive capacity of the LF at the same time using region segmentation algorithms.*

2. Define a combination schema to *improve regions discriminative capacity and improve boundary segmentation results.*

Part III

Local features discriminative capacity enhanced by region segmentation

Chapter 4

The SP-SIFT feature: Local features discriminative capacity enhanced by region segmentation algorithms

In this chapter we present the proposed method SP-SIFT. The method aims to partially validate the thesis hypothesis: local features supported by regions. First, we study the reference methods in the literature of local features combined with region segmentation algorithms. Then, we present the proposed method and exhaustively describe the resulting solution. Finally, we present the enhanced characteristics of the proposed method and conceptually validate them through experiments. This chapter is built on the work developed for the published international journal article : "SP-SIFT: Enhancing SIFT discrimination via superpixel-based foreground-background segregation", [Navarro et al., 2014b].

4.1 Method's motivation

The objective of the proposed method is to validate one of the statements presented in the thesis hypothesis. The statement claims that LF and region segmentation algorithms capabilities are complementary. Specifically, it claims that the ability of region segmentation algorithms to isolate information in the image can improve the discriminative capacity of LF.

The idea is not only to propose a specific method to validate the hypothesis. It is also to define a standard to combine region segmentation algorithms and LF.

Due to a combination of reasonable performance and publicly available implementations, scale invariant feature transform (SIFT) [Lowe, 2004] and speeded up robust features (SURF) [Bay et al., 2006] are the most popular description techniques in the LF field. These techniques share a similar relative-to-neighborhood approach for their description stage (see Chapter 2).

This approach has proven to be effective for the description of self-defined entities as images or areas inside objects. However, this approach maybe problematic for the description of objects that might be surrounded by variable backgrounds (object collections, object tracking, etc.). In such case, the description area or gradient pooling region of a local feature detected close to an object's boundary may include information from the background. As a result, it will not resemble for the same feature detected in the same object placed over a different background.

This situation is illustrated in Figure 4.1. A simplification of the SIFT detection and information extraction process is shown. In the left column, the same feature (the top part of the letter L in the sculpture) is detected by SIFT for both input images. The mid row of the first column, visually compares the description areas to be used by the SIFT description stage. The description areas to be used by the proposed method (which is described below) can be visually compared in the center of the Figure. In the right column, we can see the segmentations into superpixels that partially support the process.

As shown in the figure, the proposed solution, named superpixel-based isolation of scale invariant feature transform (SP-SIFT), is based on incorporating region segmentation into the description stage of SIFT. The proposed method is defined using the SIFT feature [Lowe, 2004], and the superpixels segmentation algorithm SLIC [Achanta et al., 2012]. However, the concept of isolating information in the local feature description stage can be performed using different region segmentation algorithms and different LF. Our results indicate that this technique achieves:

- Higher stability of the SIFT descriptor to image changes.
- Lower distance—in Euclidean terms—between local feature descriptors of objects placed over different backgrounds.

4.2 Background on local features supported with region segmentation algorithms

LF and region segmentation algorithms have been widely used in computer vision applications. However, to our knowledge, very little research has explored to combine them towards obtaining more robust computer vision tools. We describe here the closest approaches in the state of the art of this field. We also include in this background review techniques that share the motivation of our proposal, even if their proposed schema is not similar.

We propose to divide state of the art approaches in two main branches: those proposing regions as the pooling areas for the LF description process, and those proposing to include regions information in the LF description stage. We include in this section the most relevant approaches in these categories. For each method, we highlight its main characteristics and how

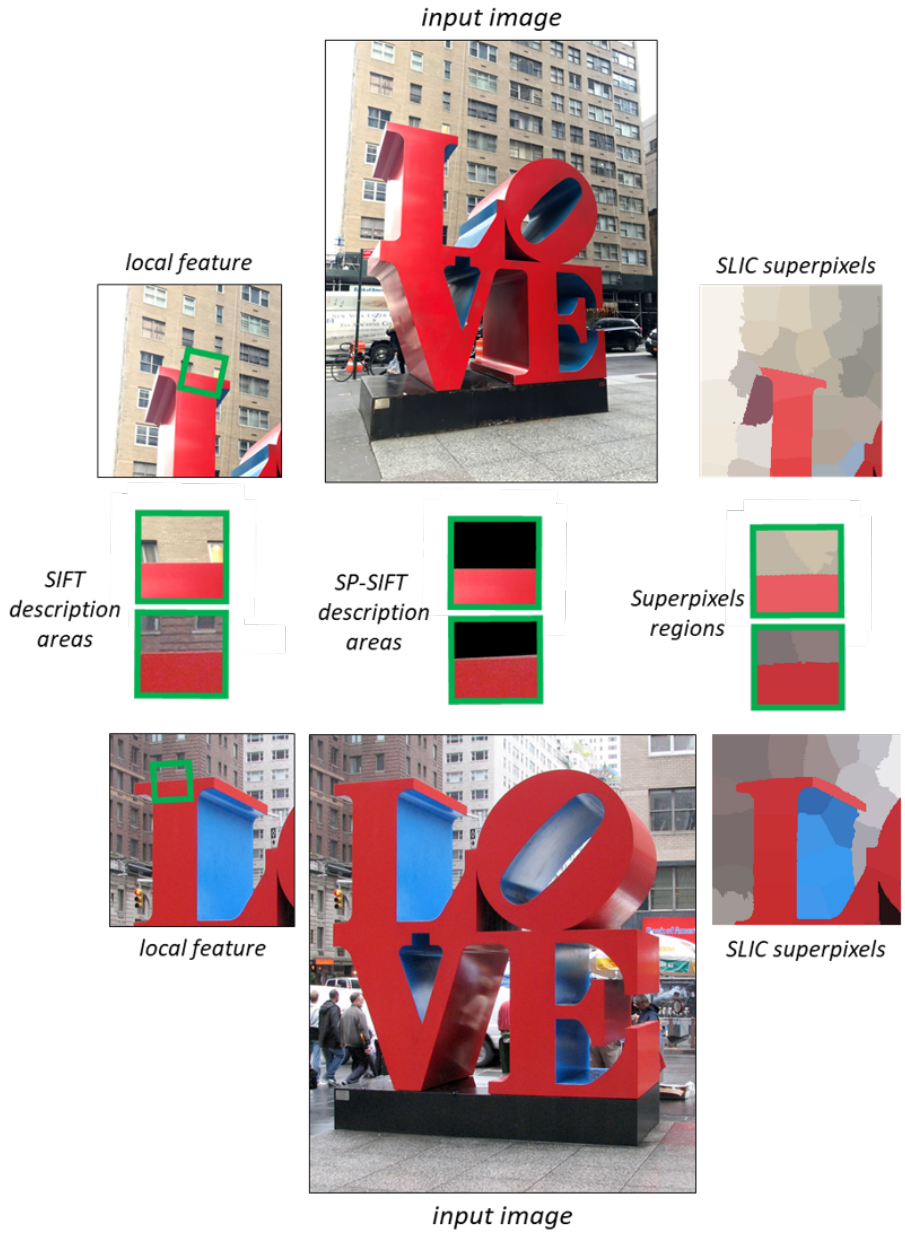


Fig. 4.1. Comparison of two LF extracted from different views of the same object. Local feature detection is similar in both images. Each image is segmented in SLIC superpixels. Mid row visually compares the information to be described by the original SIFT method and the information to be described by the SP-SIFT method. It also compares superpixel regions that appeared in the local feature description area.

they propose to combine LF and region segmentation. Further details about the techniques are not relevant for this work, and they can be found in the original papers.

4.2.1 Using region segmentation as detection support for local features description

The work described in [Koniusz and Mikolajczyk, 2009] is one of the few evaluations of region segmentation algorithms and LF available in the state-of-the-art. The article was published in a top conference (British Machine Vision Conference), and one of its authors is Krystian Mikolajczyk, probably one of the three top researchers in LF worldwide. However, it only counts 13 cites. That gives an idea of the little research that has been done in this area. The study concludes that using LF can be a good idea to measure repeatability of segmented regions, i.e. that it is possible to match regions by describing them with LF.

There are some well-known region detectors such as Maximally Stable Extremal Region (MSER) [Donoser and Bischof, 2006] that were not designed to perform as region support for LF, but further research took advantage of their capabilities in this way. There are additional region segmentation techniques including schemes that are similar to local feature techniques in their regions characterization process, e.g. intensity extrema-based region (IBR) and edge-based regions (EBR) [Tuytelaars and Van Gool, 2004], Principal curvature-based region (PCBR) [Deng et al., 2007] and Stable features [Gu et al., 2010]. Differently, there are few proposals including regions in the proper schema of local feature. The Medial Feature Detector (MFD) [Avrithis and Rapantzikos, 2011] does it but with a different objective: maximizing the detections repeatability. The FLOG method [Cui and Ngan, 2011] shares the objective of isolating information in the LF description area. However, it presents a solution based on edge detection, so all the possible information outside the edges is discarded for description.

1. MSER obtains interest regions based on thresholding the pixel intensities, i.e. it is an advance version of the watershed method. Extremal regions are defined as those in which all the pixel values are brighter or darker than those on their boundaries. Despite detecting highly repeatable regions, its boundaries are far from being stable, leading to less discriminative region descriptors.

2. IBR y EBR (intensity and edge-based regions). Like MSER, these techniques perform a reliable detection of affine invariant regions. However, they are usually described using color based techniques, which lacks of discriminative power, hindering their performance.

3. PCBR is based on MSER operating in watershed regions of principal curvature images. The principal curvature image is extracted from eigenvalues of Hessian matrices. It uses region information to improve the stability and repeatability of the detections. However, it does not provide any further improvement in the description stage.

4. Stable features share with our proposal the aim of improving the LF results using regions. However, its main contribution lies more in a post-processing fashion. Authors propose to use the information of the region segmentation to associate LF placed in the same region. Despite the improvement in the results, the local feature weaknesses are not directly addressed.

5. MFD presents a novel methodology to improve regions, stability and repeatability. It is based on a distance transform based mapping using gradient information. As a result, it obtains high repeatable regions, even without performing any sort of scale or affine variation processing. Their main contribution is the reduction of the memory and computational requirements, a major issue of SIFT like LF, which is not addressed by the proposed method.

6. FLOG proposes a solution to improve the description of edge detections. Authors introduce a scale- and affine-invariant feature called the Fan feature. It consists in a set of transformations—like SIFT main orientation normalization—that are applied to edge detections. They also propose a SIFT like descriptor of the Fan feature. They demonstrate a reasonable performance in structured scenes, where edges are the dominant characteristics. However, this method needs to be complemented with additional techniques to be suitable for more generic scenarios.

4.2.2 Using region segmentation as additional information for the local features description

Few but relevant and well-known references can be found in the state-of-the-art, most of them being SIFT like proposals. We describe here the best-known techniques as a reference for the proposed method in terms of results. More details can be found in the original papers.

1. GLOH [Mikolajczyk and Schmid, 2005] has been described in Chapter 2. It is one of the first LF that proposes to change the SIFT gradient pooling area for a circular support region. Authors also propose to use a Principal Component Analysis (PCA) to reduce the descriptor dimensionality. GLOH performs at the same level as SIFT's. We include it as a reference of successful modifications of the description area.

2. DAISY [Tola et al., 2009] has also been described in Chapter 2. We consider it a local feature, and not a region segmentation and local feature combination. It proposes an occlusion detection method to avoid including information in the description from a set of predefined occluding patterns. It is not properly based on using regions, but it shares with the proposal the idea of removing areas of the description environment.

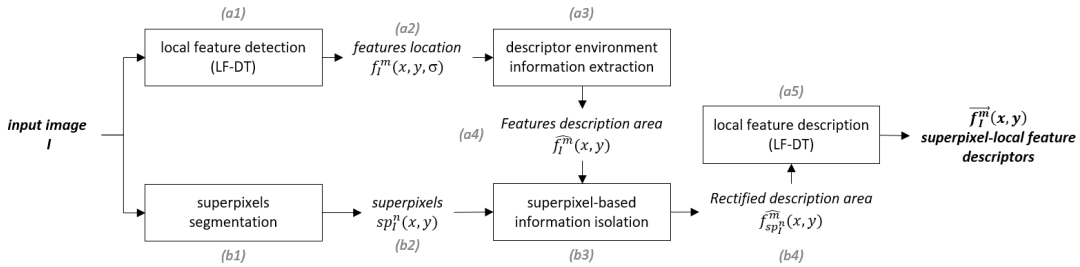


Fig. 4.2. SP-SIFT overview for a generic environment description-based local feature. LF are detected in the image (a1). In parallel, the image is segmented in superpixels (b1). Detection define the area at which the information for the description process is extracted (a3). Superpixels shapes are used to rectify the description areas (b3). The local description process is applied in the rectified areas (a5). The method results in the superpixel-local feature descriptors $\overrightarrow{f_{sp}}(x, y)$.

3. SIFT-LBP [Yuan et al., 2011] is a clear example of regions adding information in local feature descriptions to improve its capabilities. The method proposes to independently extract SIFT descriptors and local binary patterns (LBP) [Ojala et al., 1996] and integrate them in a single descriptor. It also provides a feature matching methodology to take advantage of the proposed schema.

4. Recognition using regions [Gu et al., 2009] proposes to use high quality and stable regions as a support for a fusion of descriptors. These descriptors include mainly region-based characteristics such as color, shape and texture. They do not use LF techniques but proposes a very interesting schema to merge description information.

4.3 The SP-SIFT method

Figure 4.2 shows schematic overview of the method. We will use the indexes in the Figure to guide the method description. Notice that the schema is defined for a generic local feature, with the restriction of being environment description based. The method here described is a specification of the schema using the SIFT feature and the SLIC region segmentation algorithm. Figure 4.1 will be used to graphically describe the process and compare the two LF extracted with the original SIFT technique and with the proposed SP-SIFT from a qualitative point of view.

As in the original SIFT method; the process is divided into two stages: detection and description. The detection stage is also divided in two branches that will be labeled as (a) - LF and (b) region segmentation.

4.3.1 SP-SIFT detection stage

(a1) Local feature detection process. It follows the SIFT method described in Chapter 2. For a given input image I, the method obtains a set of feature detections $\{f_I^m(x, y, \sigma)\}$, where each detection is defined by its: spatial location (x, y) and scale (σ). Whereas and $m = \{1, ..., M\}$ indexes each feature among the M. Figure 4.1 includes one detection per image, $f_{I1}^1(x_0, y_0, \sigma_0)$ and $f_{I2}^1(x_0, y_0, \sigma_0)$.

(b1) Input images $I1$ and $I2$ are segmented into SLIC superpixels. We use default parameters defined in Chapter 2. As a result, we obtain a tight to boundary segmentation. Each region (segment) is defined as the set of pixels contained in each region, $sp_I^n = (x_r, y_r)^{R_n}$, where $n = (1, ..., N)$ and N the number of regions resulting after the segmentation, $r = (1, ..., R_n)$, and R_n the number of pixels of a given region n. It is important to remark that, due to the characteristics of both methods, SIFT points are expected to be in superpixels boundaries.

4.3.2 SP-SIFT description stage

The core of SP-SIFT lies in the description stage. Let us define the concept of active area as the surface of a superpixel that overlaps with the SIFT description area. Figure 4.1 shows the candidates to be active areas for the given LF in the mid-right row, *Superpixels regions*. In order to avoid the description of size-marginal areas, active areas smaller than a quarter of the SIFT description area are discarded. Experimentally, we have observed that such a threshold represents a trade-off solution between descriptions repeatability and distinctiveness.

(a3) According to the description process defined in SIFT, a square description area is defined around each detected feature. The area is rotated according to the principal orientation of the gradient information. Figure 4.1 shows the SIFT description areas for both detected LF.

(b3) In SIFT, the descriptor and the principal orientation (used later for description normalization) are both extracted from the gradient information of the whole description area. We propose to evaluate them separately for each active area. More in detail, for every active area in the SIFT square description area, pixels of the SIFT description area laying out of the active area are removed before computing the SIFT descriptor. Figure 4.1 shows two of the SP-SIFT description areas for the detected LF.

(a5) This process results in a set of SIFT descriptors—one or several per up to four active areas—per detected SIFT point, which overall conform the proposed SP-SIFT descriptor. Each of these descriptors describes an active area. The gradient information of this modified square description area is extracted. Then, the principal orientation is computed.

Finally, a SIFT descriptor is obtained and normalized for every principal orientation, if more than one.

We do not know *a priori* which active area better describes a SIFT point. Therefore, we defined the best area as the one that minimizes the distance between SP-SIFT descriptors.

In order to evaluate the matching of SP-SIFT descriptors of two POIs, p and q, we propose the following approach:

1. Let $\overrightarrow{p_{nk}}$ be the *k-th* descriptor, $k = (1 \dots K)$, of the *n-th* active area, $n = (1 \dots N)$, of point p, where K depends of the number of principal orientations and $N \leq 4$;

2. Similarly, let $\overrightarrow{q_{n'k'}}$ with $k' = (1 \dots K')$, $n' = (1 \dots N')$ and $N' \leq 4$ be the descriptor of point q.

3. As the cardinals of the sets of descriptors for p and q might be different, we propose to evaluate the similarity between p and q as the minimum distance between their respective descriptors:

$$d(p,q) = \min_{n,n',k,k'} (\|\overrightarrow{p_{nk}} - \overrightarrow{q_{n'k'}}\|_2) \tag{4.1}$$

, where $\|\overrightarrow{x} - \overrightarrow{y}\|_2$ stands for the Euclidean distance between two vectors $\overrightarrow{x}$ and $\overrightarrow{y}$.

4.4 SP-SIFT validation test

We define two concepts that need to be tested in order to validate the proposal. Application validation will be performed in Chapter 5. First concept is to maintain or improve the general capabilities of the original method. Second concept is to evaluate the improvement obtained in the motivation-scenarios.

4.4.1 SP-SIFT capabilities validation

Capabilities validation is performed following [Mikolajczyk and Schmid, 2005]. We present briefly the objectives and the evaluation framework proposed there and how we apply it for the validation SP-SIFT.

The main goal of the evaluation framework is to identify which local feature technique perform better. Specifically, the evaluation of the local feature descriptors is performed in the context of matching and recognition of the same scene or object observed under different viewing conditions. The quality of the work, and the public availability of the dataset, have converted the framework in a reference benchmark for the evaluation of LF capabilities.

Authors propose to evaluate the descriptors on real images with different geometric and photometric transformations and for different scene types. The dataset contains six image

transformations: rotation; scale change; viewpoint change; image blur; JPEG compression; and illumination. In the case of rotation, scale change, viewpoint change, and blur, they use two different scene types. One contains structured scenes and the other contains repeated textures with different patterns. Additional details about the dataset can be found in the original paper.

The evaluation metric defined in the framework considers precision and recall. To do so, they define a correspondence criterium that depends on a nearest-neighbor distance threshold. They sweep the distance and obtain precision-recall curves. In the evaluation, they vary the detectors in order to measure its influence in the description and matching processes.

4.4.1.1 Evaluation framework specification

Objective: We do not perform an evaluation of different detectors, as the focus is on verifying the capabilities of the proposed method in the description stage.

Dataset: For the dataset, the robustness to JPEG compression was not defined as a desirable capability in Chapter 2, so it is leaved out of the evaluation. Figure 4.3 shows few samples of the dataset, a pair of images per evaluated capability. As shown in the figure, the dataset is composed of four categories or capabilities to evaluate: viewpoint, scale+rotation, blur and illumination. Each category contains six images ($L_1 - L_6$), up to 24 images. Image L_1 correspond to original image. From L_2 to L_6 the *intensity* of the property variation increases. For example, blur variations are performed using a Gaussian blur filter. The L_2-blur image is the result of applying a 0.5 pixels radius Gaussian filter, whereas L_6-blur image results from applying a 4 pixels radius Gaussian filter. Further details about the dataset can be found in the original paper [Mikolajczyk and Schmid, 2005].

Methodology-metric: We perform the evaluation in two ways. First, we evaluate the recall (repeatability) of the descriptor per capability. As we consider a one to one correspondence for the detected LF, the precision score will add no information. Additionally, we wont sweep threshold for the detection's association, but match each feature of one image to the closest unassigned one in the other image, wielding a single performance per category.

$$recall = \frac{\#correct\ matches}{\#correspondences} \tag{4.2}$$

$$precision = \frac{\#correct\ matches}{\#correct\ matches + \#false\ matches} \tag{4.3}$$

where *correct matches* are detections correctly matched and $false\ matches$ are detections wrongly matched. *correspondences* is the total number of true matches. We define a true match as two detections laying in the same position of the target. To avoid detection-related issues

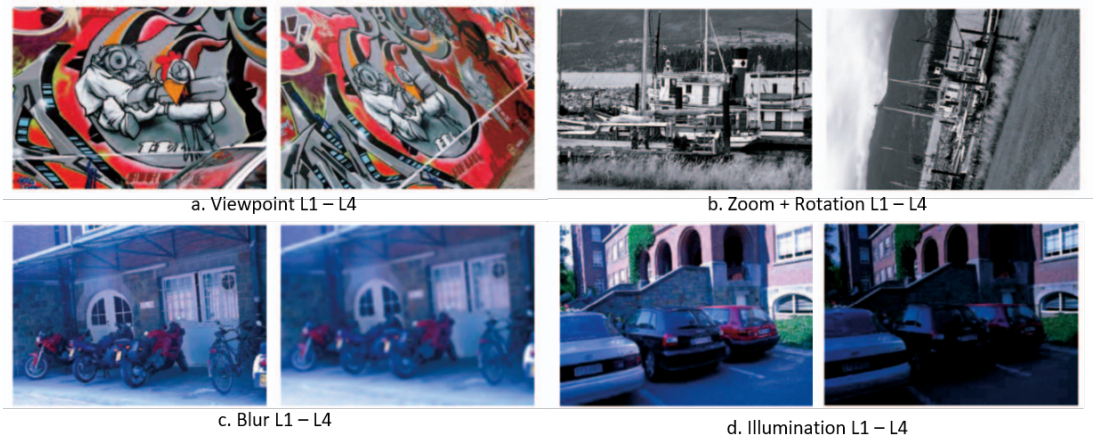


Fig. 4.3. Validation database samples. Per category, it is shown L_1 image (original) and L_4 image (4 level of variation). There are 6 levels of variation (intensity) of the original image per category.

Category	Viewpoint					Scale + Rotation				
Intensity levels comparison vs L_1	L_2	L_3	L_4	L_5	L_6	L_2	L_3	L_4	L_5	L_6
SIFT	0.47	0.37	0.22	0.07	0	0.71	0.67	0.51	0.38	0.12
SP-SIFT	0.52	0.44	0.23	0.07	0	0.79	0.72	0.55	0.40	0.12

Category	Blur					Illumination				
Intensity levels comparison vs L_1	L_2	L_3	L_4	L_5	L_6	L_2	L_3	L_4	L_5	L_6
SIFT	0.77	0.56	0.43	0.31	0.09	0.43	0.32	0.18	0.08	0.02
SP-SIFT	0.74	0.57	0.33	0.28	0.04	0.45	0.27	0.15	0.06	0

Table 4.1: For each category evaluated, recall results for LF matching between L_1 and L_X images, where $X = (2, ..., 6)$.

in the definition of a true match, ground-truth detections are defined manually. Based on this criterium, we can say that $\#correspondences = \#correct\ matches + \#false\ matches$.

Second, we evaluate the distances of the matchings. The objective here is to evaluate the potential in discriminative capacity, as in this dataset the number of distractors is very limited.

4.4.1.2 Results and discussion

Table 4.1 shows the results in terms of recall and precision for each of the four categories. We can appreciate a slight improvement of the proposal for viewpoint and scale+rotation categories. Blur and illumination are the categories where the proposal performs slightly worse than the original method. Results are not as good for these categories as the region segmentation process is highly sensitive to these changes. Despite superpixels have proven to be effective in segmenting blurred and dark images, extreme variations are still challenging.

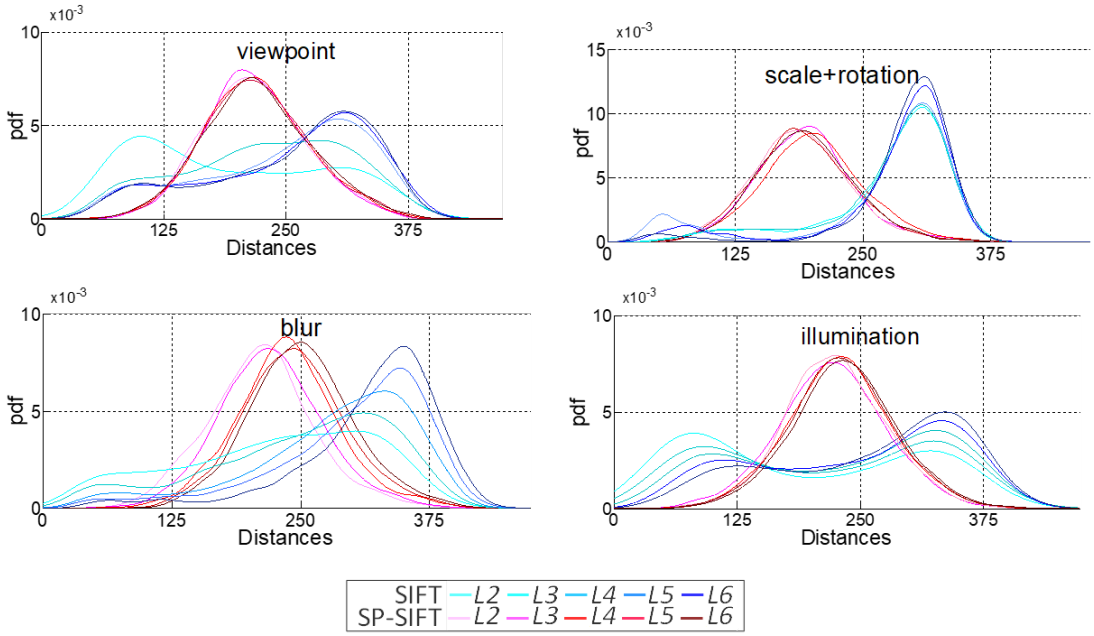


Fig. 4.4. Validation database samples. Per category, it is shown L1 image (original) and L4 image (4 level of variation). There are 6 levels of variation (intensity) of the original image per category.

Results allows to conclude that the proposal presents the desired capabilities of the original method.

Figure 4.4 shows four graphics, one per category in the dataset. They represent the probability distribution function (pdf) of the *correct matches* matching distance per image $L_2 - L_6$, and per technique. SIFT matching distances in all the categories are distributed in the whole range. SP-SIFT matching distances tend to be more concentrated. There are a number of SIFT matching on every category with distances lower than SP-SIFT's. However, there also a number of SIFT matching distances on every category with associated distances larger than SP-SIFT's. SP-SIFT seems to be more robust to image variations, as its distance distribution remains almost unaffected by complexity. As a conclusion, the original proposal provides more discriminative descriptors, it tends to use more information in the description process. However, in overall, the matching distances of the proposal are lower than the original technique; hence, more discriminative.

4.4.2 SP-SIFT foreground-background segregation validation

4.4.2.1 Evaluation framework specification

Objective: The objective of this validation test is to measure the improvement of the proposal in the main objective, i.e. the capability of describing a target disregarding environment

Fig. 4.5. Foreground-background segregation validation database example. Isolated targets are overlaid on a set of 4 textured background.

information variation. The best example of this task is a tracking application. These application validations will be performed in Chapter 5. In this section, we validate the proposal defining a target-oriented evaluation framework.

Dataset: We define an experiment-oriented database. We use as baseline the proposed database [Tiburzi et al., 2008]. The original database was recorded using a chroma key, so target objects can be isolated from the background. We built the dataset taking the real objects and overlaying them over four highly textured backgrounds. The original database is a video database. We selected 3 frames of each sequence in the database (9), up to 27 foreground object images. Extracted foregrounds are overlaid the four textured backgrounds for a total of 108 images.

Methodology: Every so-built image is compared against each other (per target object), adding up to six comparisons. The experiment is two-fold: first we evaluate the ability of each descriptor in matching all objects LF. Then, we focus only in the matching of objects boundary detections, which are affected by foreground-background effects. We threshold the matching distance and account for the average precision and recall curves for the dataset.

4.4.2.2 Results

Results are illustrated in Figure 4.6, via a classical precision recall study by thresholding the matching distance. In the light of these curves, SP-SIFT outperforms SIFT in both experiments. In the task of objects description, SIFT generally yields to lower distances than SP-SIFT when the LF detections are full-contained inside the objects (reflected also in the lower modes in Figure 4.4). However, SP-SIFT better discriminates these LF detections respect to background detections, then ranking equally (or better) at objects-inside LF. The main differences between SIFT and SP-SIFT arise in boundary points: SIFT's description of these points hinders its overall operation as they include background information, whereas SP-SIFT adequately isolates the

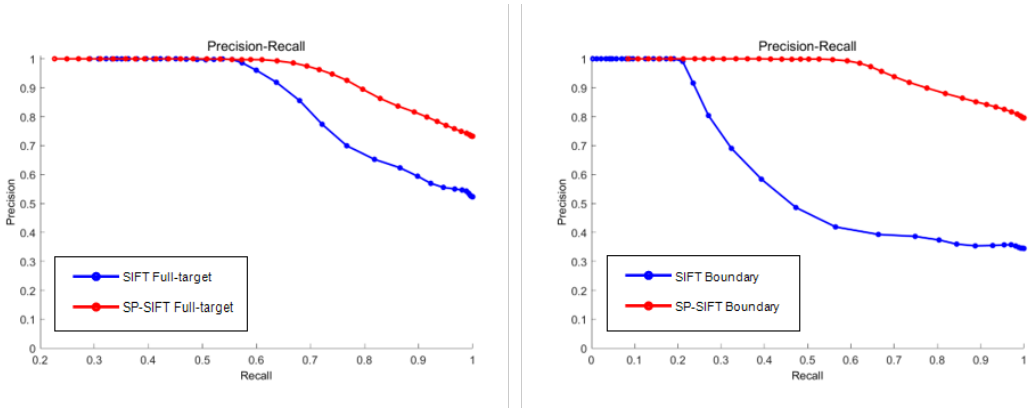


Fig. 4.6. Precision recall curves. Per category, it is shown L1 image (original) and L4 image (4 level of variation). There are 6 levels of variation (intensity) of the original image per category.

objects information. Although this test may only be considered as a validation tests, it reasonably shows the advantages of describing with SP-SIFT descriptor, especially in the description of boundary points.

4.5 Conclusions

The proposed combination of LF and region segmentation, illustrated through the SP-SIFT method, overcomes SIFT's limitations on scenarios where the description of the object of interest is disturbed by the surrounding information. This is achieved using tight-to-object superpixels that drive the isolation of the objects parts in the description and allow its reorganization. The benefits of SP-SIFT in terms of description stability and discriminability are shown in two experiments but will be extended in the following chapter in real scenario applications. Essentially, this chapter proposes a method to partially validate the thesis main hypothesis.

Chapter 5

SP-SIFT validation via tracking applications

In this chapter we present two proposed applications for the SP-SIFT method. We motivate the application selection. We present a first application where the method works as a supporting technique. We present also a second application where the method is the core solution. As a result from the applications evaluations we validate the proposed technique and thus the first part of the thesis main hypothesis. This chapter is built on the work developed for the published international conference article: "Enhancing region-based object tracking with the SP-SIFT feature", [Navarro et al., 2014a], and the pending-to-be-published article: "HPSTr: Homography Point-based Shape-fitted Tracker", [Navarro et al., 2018b].

5.1 Applications selection and integration

The proposed solution, SP-SIFT, have been validated from a functional point of view in Chapter 4. However, it is in real conditions testing where the solutions, and thus the hypothesis, can be considered proven. The motivations of the SP-SIFT feature from the capabilities point of view are clear. However, in the application level the objective is the following: mitigate the LF weaknesses so we can take advantage of their properties in new applications.

To select the applications we search for: 1) applications where LF are rarely used because of their weaknesses; 2) extensively researched applications, so we can compare our proposal with top quality algorithms; and 3) impactful applications, where the improvement can be seized and future researchers can advance on our work.

After considering a number of applications, we selected **object tracking** because of the following reasons:

- Challenge: object tracking can be defined as the task of associating a target position in

time. This definition implies the task of identifying a target in consecutive frames over varying backgrounds. This is exactly the challenge that our proposal was intending to solve at LF level.

- Researched application: object tracking is possibly the most extensively studied task in the computer vision research community. One can obtain almost 18.000 references in Google Scholar just for 2019. Every year there is a number of international challenges, special issues in top journals and surveys compiling the most recent and successful proposals.
- Impact: object tracking is one of the basic tools in computer vision. Its applications are limitless, from surveillance to customer service, autonomous driving or maintenance tasks.

Defined the task, we proposed a two steps application process.

1) SP-SIFT as a complementary technique for tracking, Section 5.2. We identify a contrasted tracking solution where our proposal can be useful. We integrate the feature in the tracking process and evaluate the level of improvement. In this way, we can quantify the net improvement, disregarding the tracking method's effect.

2) SP-SIFT as the core of a tracking algorithm, Section 5.3. We propose a tracking solution based on the SP-SIFT feature. Despite the challenging process of creating an algorithm to compete with the state of the art, this strategy allow us to validate the real potential of the proposed method.

5.2 SP-SIFT as a supporting technique: the SP-SIFT tracker

5.2.1 Region based trackers background and integration strategy motivation

In the light of the results of a number of several approaches [Wang et al., 2011; Lu and Hager, 2007; Ren and Malik, 2007; Oron et al., 2015] mid-level visual cues—regions, patches, superpixels—have proven to be effective cues for tracking. High-level appearance models [Kwon and Lee, 2010; Santner et al., 2010] are prone to contain descriptive information of the target structure, which is a must when facing intensive occlusive situations and sudden target motions, albeit they tend to be less effective in handling non-rigid target distortions. Low-level cues, here defined as LF, [Ren and Malik, 2003; Ta et al., 2009], on the contrary, adequately adapt to non-rigid distortions; however, they are less accurate in handling heavy occlusion and clutter. Mid-level cues share with low-level cues their ability to adapt to non-rigid distortions whereas they are also robust to occlusions as high-level models.

On the other hand, as exposed in [Smeulders et al., 2013], most of the tracking algorithms can be roughly divided in two main groups attending to their tracking strategy: direct matching [Nguyen and Smeulders, 2004; Comaniciu et al., 2000; Baker and Matthews, 2004] and discriminative target-background approaches [Kalal et al., 2010; Wang et al., 2011]. Matching strategies

rely on trying to maximize the direct match between a model of the target and the incoming image, whereas discriminative target-background methods use a classifier to distinguish the object by maximizing on the discrimination of the target from the background.

The results of a tracking method are obviously implementation-dependent but they are strongly influenced by the chosen strategy. Classical tracking methods [Nguyen and Smeulders, 2004; Comaniciu et al., 2000; Baker and Matthews, 2004] are usually based on matching algorithms, whereas recently, trackers built on a discriminative strategy are preferred, as they produce better results [Smeulders et al., 2013] in most of the challenging tracking situations: object drifting, sudden light changes, surface cover—or appearance—and size-shape target changes.

With this in mind, it seems natural to include mid-level visual cues in a tracking method based on a discriminative strategy. However, the success of this configuration is biased by an inaccurate characterization of the mid-level cues. Discriminative strategies in other areas of computer vision as object recognition and classification, usually rely on LF-DS [Lowe, 1999; Tola et al., 2008; Mikolajczyk and Schmid, 2005] achieving robust and accurate results. On the contrary, in object tracking, mid-level cues are usually described by less-discriminative cues as region color average or luminance and color histograms.

We aim to combine the advantages of 1) accurate local feature descriptors, 2) robust mid-level cues and 3) flexibility of discriminative approaches, in a tracking scheme. The proposed technique, SP-SIFT, has the capacity of describing mid-level cues—in this case SLIC superpixels, SP—using LF description schemes—in this case the Scale Invariant Feature Transform, SIFT. In the current application, we proposed to include this feature in a state-of-the-art discriminative tracker based on mid-level cues, in order to measure the effect of describing these cues with LF-DS, and evaluate the advantages performed by the proposed method. Furthermore, the current SP-SIFT application aims to pave the way to extrapolate LF-DS to mid-level cues avoiding to include noisy information but maintaining the desired properties from classical LF methods.

5.2.2 Baseline tracking algorithm

We take into account the conditions mentioned in Section 5.1. The objectives are to work with a state-of-the-art object tracker and evaluate the claims made about the SP-SIFT feature. The selected initial method should fulfill the following conditions:

1. It should follow a discriminative target-background strategy.

2. It should use mid-level cues, preferably superpixels, because it is the cue used in the proposed feature algorithm.

3. It should be based on some of the trending cues for mid-level cues description, as HSI-histograms (hue,saturation and intensity histogram).

With this in mind, we selected a well-contrasted tracking algorithm well-ranked in lately surveys, which we briefly describe in the following lines. Further descriptions can be found in the original article.

The method, called superpixels tracking (SPT) [Wang et al., 2011], is a high precision tracker based on the effective and efficient use of superpixels (condition 2) as mid-level visual cues. During a previous training stage, the segmented superpixels are clustered to construct a discriminative appearance model (condition 1) via assigning clusters either to target superpixels or to background ones. The description feature to construct the appearance model is the HSI-histogram of each superpixel (condition 3). In the test stage, a confidence map at superpixels level based on their HSI-histogram is computed, using the appearance model to obtain the most likely target location with maximum a posteriori (MAP) estimates. The appearance model is constantly updated to account for variations caused by change in both the target and the background. Experimental results on various sequences [Wang et al., 2011] show that the selected algorithm performs favorably against alternative state-of-the-art methods. The algorithm also includes an occlusion-avoiding technique, as well as motion and scale prediction.

The strategy followed to develop the model has proven to be effective, whereas the distinctiveness of the resulting model is less remarkable. The low distinctiveness of HSI-histogram may cause superpixels to wrongly move between close clusters on the HSI-histogram space even when they belong to different target-background category.

Even if the initialization and classification of the superpixels into target or background clusters is correct, the generation of the confidence map depends on the discriminative capacity of the HSI-histogram feature. This capacity can be considered a weakness since any two regions with similar average values of hue, saturation and intensity are recognized as akin, even if their visual information is completely different. Finally, despite the use of confidence maps at superpixel level, the target region of the SPT algorithm for a given frame is a bounding box. Although a bounding box is the most used target shape in order to test tracking algorithms, evidence suggests that results at mid-level define much more precisely the target shape. The selected algorithm is not able to give results for each target region due to the use of HSI-histograms. Histograms cannot capture any spatial ordering, which should be captured elsewhere in the tracking algorithm. That is the reason why a confidence map and its bounding box is needed to locate the target.

5.2.3 Integration of the SP-SIFT feature

The selection in the SPT tracker of mid-level cues in combination with a discriminative target-background strategy should handle almost every kind of target self-event, i.e., shape and motion distortion or appearance changes. Additionally, the occlusions-avoiding and the motion and scale prediction procedures should make this algorithm robust against the interactions of the

scene elements with the target and against camera distortions. In conclusion, it should be able to overcome almost every difficulty in video tracking.

Despite this conclusion, the evaluation in [Smeulders et al., 2013] reported that when the algorithm was tested without pre-adapting manually the initial parameters to the scene conditions the results of the tracker were not as expected. Main reported issues were related to background distractors. The proposed inclusion of the SP-SIFT feature into the selected tracker is expected to solve or at least mitigate these problems related to the low distinctiveness and discriminative capacity of the mid-level cues descriptors used. This should also lead to a less-dependent algorithm from its initial parameters. In this direction, the simpler way to include SP-SIFT requires modifications in two stages: the discriminative model and the confidence map generation.

5.2.3.1 Enhancing the discriminativeness of the target-background model

As aforementioned, the selected algorithm constructs the target-background model during a training (T) stage. Superpixels are characterized by their HSI-histogram and then clustered. Finally, each cluster is assigned to either the target or the background according to an overlapping measure respect to the ground truth bounding box. Now, let $F^T = \{f_1, .., f_n, .., f_N\}$ be the set of N SP-SIFT LF that we in parallel extract for the detected LF in the frames of this training stage, each feature associated to the superpixel that describes; and let $F_i^T \subset F^T$ be the subset of N_i LF corresponding to the i^{th} cluster, which we include in the target-background model complementing with LF the distinctiveness provided by HSI-histograms. We finally obtain for each cluster a coherence or matching measure among all the SP-SIFT LF belonging to that cluster as:

$$M_i = \frac{\sum_{l=1}^{N_i} \sum_{m=1}^{N_i} \|f_l - f_m\|_2}{\binom{N_i}{2}} \tag{5.1}$$

5.2.3.2 Confidence map generation

In the test or operation (O) stage of the selected tracker, every incoming frame is segmented into P superpixels and the HSI-histogram descriptors $H^O = \{h_1, ..., h_p, ..., h_P\}$ are extracted for them. Then, the original algorithm obtains a confidence map for these P superpixels deciding to which model cluster they match, and evaluating the confidence of such matching. This results in subsets $H_i^O \subset H^O$ of superpixel descriptors matching each to the i^{th} cluster.In parallel, we extract $F^O = \{f_1, ..., f_l, ..., f_L\}$ SP-SIFT LF for the same frame. We then group these LF so that $F_i^O \subset F$ is the subset of LF belonging to the superpixels characterized by the H_i^O descriptors. Observe that while every superpixel will have a h_p descriptor associated, it might have one, several or no SP-SIFT LF associated, depending on the number of LF detected in such superpixel.

A matching of SP-SIFT LF is then performed to assign every feature of the set $F_i^O$with LF of the set F^T, which LF were associated to some of the clusters from the training stage.

Each of the LF from the set F_i^O are then matched to one of the trained clusters classified as target or background. As a result of this matching stage, and avoiding to take into account matchings with an associated distance higher to its correspondent M_j, being j the cluster to which they matched, the F_i^Oset has S LF associated with target-classified clusters and $\left|F_i^O\right| - S$ LF associated with background-classified clusters. From this *voting* process, all the superpixels associated to the i^{th} cluster on the histogram matching are weighted as the result of the voting, i.e. the confidence map is refined including the new confidence of being target or background depending of which has been voted as the most probable. Superpixels with its own SP-SIFT descriptor defines its confidence just depending on its descriptor matching, and not on the voting result or the confidence map. This strategy allows to avoid residual HSI-histogram matchings due to color similarities and low discriminative capacity of the histograms. A visual example is presented in Figure 5.1.

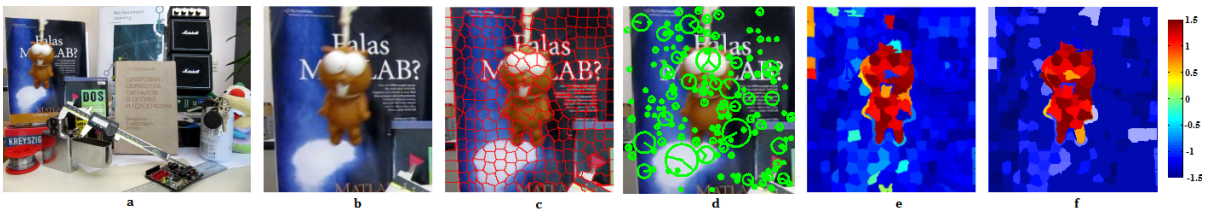


Fig. 5.1. Stages of SPT SP-SIFT a. Current frame b. Estimation of the searching area c. SLIC superpixels segmentation d. SP-SIFT detections e. HSI-histogram based confidence map (SPT results) f. SP-SIFT confidence map refinement (proposal). Note how the confidence map mismatches of SPT are corrected by means of SP-SIFT.

5.2.4 Experimental results

5.2.4.1 Evaluation framework

In order to fairly assess the advantages of the proposed modification, we have extracted the experimental results over the same sequences used in the original paper [Wang et al., 2011] describing the SPT tracker. We have also first evaluated our proposal with some sequences used in previous works: "singer1" and "basketball" from VTD [Kwon and Lee, 2010], "transformer" from PDAT [Kwon and Lee, 2009], "lemming" and "liquor" from PROST [Santner et al., 2010], and "woman" from Frag [Adam et al., 2006]; and then we tested four sequences from the SPT author dataset: "bolt", "bird1", "bird2", and "girl". For a broader comparison, we also reproduce results for the trackers that SPT compared to in its evaluation [Wang et al., 2011].

To test the proposed modification under the same conditions of the original article, the parameters of the tracker have been set to the same values as proposed by the author for each

Sequences\Methods	MS	PF	IVT	Frag	MIL	PROST	VTD	SPT	SPT SP-SIFT
lemming	236	184	14	84	14	23	98	7	**3**
liquor	137	28	296	31	165	22	155	9	**5**
singer	116	25	5	21	20	-	**3**	4	5
basketball	203	21	120	14	104	-	11	**6**	10
woman	32	76	133	112	120	-	109	9	**8**
transformer	46	46	131	47	33	-	43	13	**12**
bolt	204	34	386	100	380	-	14	6	**4**
bird1	330	137	230	228	270	-	251	**15**	**15**
bird2	73	75	115	24	13	-	46	**11**	12
girl	304	16	184	106	55	-	57	21	**15**
average	*168,1*	*64,2*	*161,4*	*76,7*	*117,4*	*22,5*	*78,7*	*10,1*	***8,9***

Table 5.1: Tracking precision results. The numbers denote the average error in the location in pixels of the bounding-box center.

sequence.

The results have been extracted in two ways to show both the precision of the proposal and its robustness. A hard metric has been used to check the tracking precision: values for the average error in the location in pixels of the bounding-box center have been obtained. In order to evaluate the robustness, i.e.e the capacity of not losing the target along the sequence, we have used the evaluation metric of the PASCAL VOC. Due to the fact that the discrimination against clutter is one of the hardest challenges for HSI cues, a frame by frame precision graphic on the "lemming" sequence has been additionally extracted to prove the capacity of the SP-SIFT feature to overcome this disadvantage of the SPT tracker.

5.2.4.2 Experiments

Precision: The average improvement respect to the base tracker algorithm reaches the 12 % with the inclusion of the SP-SIFT feature, as it is shown in Table 5.1. Despite the fact that the SPT tracker results are remarkable in this evaluation, the inclusion of a more discriminative feature on the process results in a superior performance in 6 out of 10 sequences. Outstanding results can be observed for instance in "lemming" or "bolt" sequences, and higher improvement is achieved in "girl" sequence and again in "lemming" sequence, in the latter due to the presence of clutter and color appearance similarities between the environment and the target. Qualitative example is included in Figure 5.1.

Robustness: The robustness of the base tracking algorithm was reported in the original paper, where it overcame all the compared state-of-the art trackers in all the sequences, except for the "singer" one, where the colors of the target are very similar to the background during several frames. In the results obtained for the proposed modification (see Table 5.2), a slight but remarkable improvement is obtained, enhancing the base tracker performance in 7 out of 10

Sequences\Methods	MS	PF	IVT	Frag	MIL	PROST	VTD	SPT	SPT SP-SIFT
lemming	171	426	1046	678	1105	969	471	1290	**1316**
liquor	413	1202	380	1375	353	1444	471	1701	**1720**
singer	64	96	332	84	84	-	**350**	297	296
basketball	78	455	80	512	204	-	601	**707**	695
woman	35	31	49	44	38	-	27	310	**311**
transformer	28	32	29	38	30	-	47	124	**120**
bolt	15	172	4	32	12	-	195	224	**230**
bird1	1	6	4	47	114	-	7	139	**140**
bird2	36	19	9	42	86	-	9	**94**	91
girl	79	1106	107	628	560	-	828	1180	**1298**
average	*92*	*354,5*	*204*	*348,3*	*258,6*	*1207*	*300,6*	*606,6*	***621,7***

Table 5.2: Tracking recall results. The numbers denote the count of successful frame based on evaluation metric of the PASCAL VOC object detection [Everingham et al., 2010].

sequences. The notorious results achieved for sequences "lemming", "bolt" and "girl", due to aforementioned reasons, are confirmed in this experiments.

5.2.5 Discussion of the SP-SIFT supporting application experiment

Proposed improvements due to the description of mid-level cues used in discriminative tracking schemes have been proven in this section. The inclusion of the SP-SIFT feature in the SPT tracker has allowed to demonstrated how LF techniques can improve even top-ranked state-of-the art trackers results by providing them more discriminative and distinctive capacity. This results pave the road to including LF in other kind of mid-level cues based applications, guaranteeing an increment on its discriminative and distinctive capacity. This results also prove that the SP-SIFT method present the foreground-background segregation capability presented in Chapter 6, thus partially validating the thesis hypothesis.

5.3 SP-SIFT as the main technique: HPSTr: Homography Point-based Shape-fitted Tracker

5.3.1 Object tracking background

The main objective of single-object tracking approaches—from now on trackers—is to detect the position of the target—the tracked object—in each frame of a video. As demonstrated in the previous section, the accomplishment of this task depends on to the success of the tracker in facing a set of *challenges*, inherent to video sequences. We define these challenges as categories used to explain and arrange the potential changes that affect the appearance of both the target and its context—understood here as the rest of the scene. According to [Smeulders et al., 2014;

Wu et al., 2013; LIRIS, 2014], most common challenges are:

- Target-related: scale, structural and color-appearance changes;
- Scene-related challenges: clutter, inter-object occlusions, global and local illumination changes;
- Capture-related challenges: noise and camera gain.

To cope with these challenges, a substantial effort has been devoted to the design of models that adapt to the target appearance in each frame. We can define a category of algorithms proposing different updating mechanisms as online learning trackers (OLT) [Babenko et al., 2009; Kalal et al., 2010; Lim et al., 2004]. OLT methods generally improve the performance of non-learning state-of-the-art tracking approaches as shown in evaluation surveys [Smeulders et al., 2014; Wu et al., 2013] and tracking contests [LIRIS, 2014].

In OLT approaches, the position of the target in each frame is usually provided in the shape of a of a bounding-box (BB), defined as a rectangle enclosing the target. Nowadays, BB can be considered the paradigm on which to build evaluation measures and hence, the expected tracker output. In particular, the quality of the tracker is usually evaluated [Čehovin et al., 2015] via the overlap Φ—the average overlapping area between the output BB and a human-annotated BB along the video—and the failure rate F_τ—the number of frames the tracker looses the target. A failure is usually declared if the overlapping area, Φ, falls below a fixed threshold, e.g., 50% overlap in the PASCAL VOC [Everingham et al., 2010]. This metric was previously used in Section 7.2.

BB-based detections usually cover a wide area around the target. Therefore, its use generally withstands high overlap thresholds not always correlated with a *good* tracking operation. Figure 5.2 shows an example of this issue: the Φ measure extracted for the detected BB and the target mask intersection (or intersection-over-union) are compared. In this case, the object is just partially covered by the BB (8% of $\cap/\cup$) but, instead, the value of Φ indicates a proper detection (51%).

BB-outputs show additional drawbacks associated to the output format. Due to its rough definition of the target, they tend to incorporate background—and occluders—information in the target description. This information may lead to the perturbation of the target's model. Remarkable efforts have been done to cope with these situations, e.g. learning only the most discriminative cues of the target [Collins et al., 2005]. However, this may fall into collateral problems, as scene distractors in cluttered backgrounds become a challenge if the model is too general [Duffner and Garcia, 2013]. Overall, BB-based outputs provide information on the position of the target in the frame, but not necessarily on its spatial extent nor on its shape.

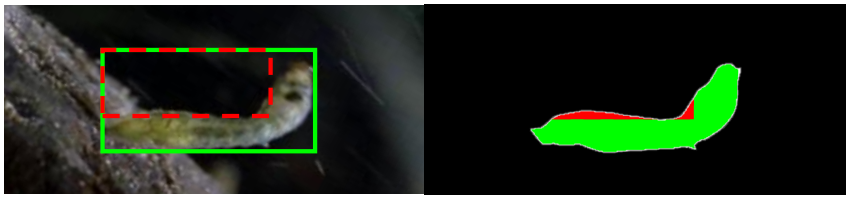


Fig. 5.2. From left to right: results at BB level and its associated measure; and object level and its associated metric, intersection over union. In red, algorithm detection; in green, ground-truth annotation.

Fig. 5.3. From left to right: detail of the *soldier* sequence Li et al. [2013], BB ideal output and OTS ideal output.

As opposed to BB approaches, Object Tracking by Segmentation (OTS) yields a binary mask (BM) clipped to the target silhouette—or a set of pixels composing this mask. See a comparison of both outputs in Figure 5.3.

The use of BMs prevents the inclusion of not-target information in the model learning process, thus decreasing the influence of drifting and occlusions. Additionally, as the target model requires a lower level of generalization, OTS trackers can reasonably cope with challenges as camouflage and clutter. Finally, tracking results satisfy the requirements of applications demanding a pixel-level definition of the target.

Nevertheless, OTS trackers present also some major challenges. Classical tracking challenges, like bouncing motion or appearance changes, successfully coped by state of the art BB-based approaches are still problematic to OTS trackers. Additionally, OTS trackers are not yet able to handle uncovered foreground situations. Besides, many of the state the art OTS approaches [Wen et al., 2015] are defined as supervised algorithms that operate offline, by analyzing the complete video sequence. For all these reasons, OTS trackers do not usually rank top in the tracking contests [LIRIS, 2014; Smeulders et al., 2014] and are discarded when analyzing long-term scenarios.

5.3.2 Proposal motivation and related work

This section describes an unsupervised online OTS tracking approach. Our aim is to provide OTS-like tight-to-object output while ensuring OLT-like robustness to tracking challenges.

To this aim, we rely on combining tracking properties of the proposed schema for LF and regions—SP-SIFT. Specifically, the contributions are three-fold:

- A scheme to propagate evidences along frames by part-based homographies obtained from the matching of robust-for-tracking LF—SP-SIFT.

- A discriminative strategy to perform BM-level object reconstruction on each frame using superpixels descriptions.

- A frame-by-frame adaptive algorithm to guarantee long-term performance.

The objective of this Chapter is two-fold: propose a state-of-the-art tracking solution and measure how far the SP-SIFT schema can wide the LF application scope.

Keeping in mind the target contributions, the tracking state-of-the art is reviewed according to the following criteria:

- Spatial arrangement of the results. We distinguish between trackers returning a bounding box (BB) and those returning a binary mask (BM).

- Temporal arrangement of the results. Although trackers generally operate returning the target position for every input frame (i.e., online operation), some strategies—mainly focused on video segmentation and edition—require the analysis of several frames or even of the whole sequence, i.e., offline operation.

- Tracking strategy. Matching and discriminative strategies are commonly used to arrange existing approaches [Smeulders et al., 2014].

- Model updating. We here differentiate between the trackers that learn or update the target or scene model as they operate (i.e., learning-approaches) and those that use the same model throughout the sequence (i.e., non-learning ones).

- Supervision. Tracking algorithms may require a quite complex parametric setup to perform on different scenarios. According to this requirement, we classify them into supervised (*ad-hoc* parametrization for each sequence) and unsupervised (a common parametrization for a data-set).

5.3.2.1 Trackers returning bounding-box results (BB)

To evaluate the hypothetical benefits of our proposal, in Section 5.3.4 we compare it against several state-of-the-art BB-based trackers. Non-learning matching-based approaches have dominated the tracking area during the last decade. Most relevant approaches in this vein include MS [Collins, 2003], PF[Nummiaro et al., 2003] and Frag [Adam et al., 2006]. Learning, but non deep learning, upgrades of these matching trackers—IVT[Lim et al., 2004], PROST [Santner et al., 2010] or VTD [Kwon and Lee, 2010]—outperformed them by including strategies to update the target model.

According to [Smeulders et al., 2014], discriminative approaches are standing out recently as better solutions to complex tracking challenges. Among these trackers, we can distinguish between those just defining a target model in order to discriminate the target from the rest of the scene—the background—and those that maintain also a background model to enhance discriminability. The former including the unsupervised and learning methods described—TLD [Kalal et al., 2010], Struck [Hare et al., 2011] and MIL [Babenko et al., 2009]—and also non-learning examples in this category but they are considered less relevant. The latter category include supervised learning trackers as SPTrack [Wang et al., 2011] and R-SPTrack [Yang et al., 2014].

5.3.2.2 Trackers returning binary mask results (BM)

BM-based trackers need to cope with segmentation-related challenges (e.g. foreground-background camouflage and clutter). The datasets proposed in [Tsai et al., 2012] and [Li et al., 2013] are considered a reference for the evaluation of these trackers. They include representative examples of these segmentation challenges, but lack of classical tracking challenges. This situation has motivated the use of foreground segmentation techniques for offline tracking. Examples of these techniques are KeySeg [Lee et al., 2011], EGraph [Grundmann et al., 2010], or [Zhang et al., 2013]. We understand these approaches as video segmentation algorithms (VOS) not as tracking algorithms.

Among BM methods that perform online (OTS), SPT [Li et al., 2013] represents one of the few unsupervised approaches. An upgraded version of this tracker (SPT+CSI) is described in [Li et al., 2013]. It improves SPT operation in cluttered environments by including a refinement stage based on region segmentation, at the expense of operating offline.

Up to our knowledge, and as discussed in [Wen et al., 2015], the rest of relevant OTS approaches are supervised. This is mainly due to their use of complex segmentation-based techniques. For instance, DynGraph [Cai et al., 2014] relies on superpixels segmentation to handle challenges as occlusions and target deformations. HB [Godec et al., 2013] integrates a Hough forest classifier together with a segmentation based on Gaussian mixture model and a graph cutting via max-flow/min-cut optimization. All these approaches include segmentation

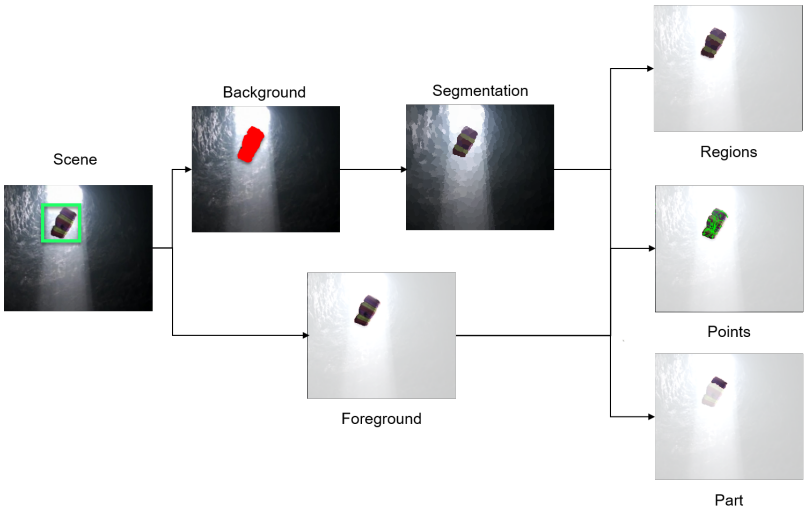


Fig. 5.4. Scene semantic hierarchy. From left to right, the scene is splitted in target and background. The background is divided in regions. The target is divided into parts, each containing regions and feature points.

stages which are strongly dependent on the analyzed scene, thus requiring supervision (*ad-hoc* configuration) to properly operate.

The JOTS algorithm [Wen et al., 2015] achieves top-performance in terms of intersection-over-union among state-of-the-art OTS on the reference data-set [Li et al., 2013]. As an online but supervised approach, it is specially tuned to segmentation-oriented evaluations. However, JOTS operates under a hand-craft parametrization. Furthermore, it operates at pixel level, albeit pixel-level has been recently proven [Yang et al., 2014] to worsen region-level in tracking challenges. As aforementioned, this approach achieves the best segmentation results for the evaluation datasets, but tracking results are quite poor, as shown in evaluation section.

The tracker proposed in this section falls in the category of BM-trackers and follows a discriminative strategy. It provides results for each video frame—under an online premise—, does not require *ad-hoc* parametric setting and relies on a learning strategy to update the target model on each frame.

5.3.3 HPSTr description

Our proposal relies on a semantic hierarchy of the scene to analyze, see Figure 5.4. The scene is divided into a target—the element to track—and the background—the information surrounding the target. The background is partitioned into regions. The target is divided into parts defined as non-deformable areas of the target. Each object's part is further described by regions and feature points.

5.3.3.1 Problem formulation

The background $\mathcal{B}$, is divided in a set of regions $\left\{r_k^{\mathcal{B}}, k = 1...K\right\}$, with K the total number of regions in the background.

In a similar way, the target $\mathcal{T}$, is divided in a set of parts $\{\Omega_j, j = 1...J\}$, being J the number of parts. Likewise, each part Ω_j is composed of K_j regions $\left\{r_{j,k}^{\mathcal{T}}, k = 1...K_j\right\}$. Together with the regions, each part is also described by a set of LF.

Regions are extracted following [Achanta et al., 2012] and described by a 5-dimensional vector containing the mode of the distribution of the region RGB values and the position of the region centroid. Thereby obtaining a feature vector describing each background region, $\mathbf{f}(r_k^{\mathcal{B}})$, and a feature vector for each region in a target's part, $\mathbf{f}(r_{j,k}^{\mathcal{T}})$, which compose the set of regions of the frame, $\mathbf{f}(\Psi_t)$

To define target's parts, we group regions via a SLIC merging approach proposed in Dollár. The algorithms is improved in our proposal.

LF are obtained using [Navarro et al., 2014b] which has been proven useful against occlusions and clutter [Navarro et al., 2014a]. The description of each target's part in terms of the N_j LF extracted on each part is composed of a set of 128-dimensional description vectors $\gamma_n(\Omega_j)$, one per detected feature in the spatial support of the part, $\mathbf{g}(\Omega_j) = \{\gamma_n(\Omega_j), n = 1...N_j\}$, being the complete set of LF of the frame defined under the notation $\mathbf{g}(\Psi_t)$.

Let $\mathcal{M}(\Omega_j)$ be a BM for each part defined as the union of the target regions spatial supports $\mathcal{M}\left(r_{j,k}^{\mathcal{T}}\right)$ composing a part, $\mathcal{M}(\Omega_j) = \bigcup_{k=1}^{K_j} \mathcal{M}\left(r_{j,k}^{\mathcal{T}}\right)$. The final output of the algorithm, a BM $\mathcal{M}(\Psi_t)$, is defined as the union of the $\mathcal{M}(\Omega_j)$, $\mathcal{M}(\Psi_t) = \bigcup_{j=1}^{J} \mathcal{M}(\Omega_j)$, where Ψ_t is the frame under analysis.

The strategy is to track a target as a set of parts. Assuming that the target parts represents planar areas of the target, the motion associated to each part, Ω_j, between consecutive frames is defined via an homography matrix $\mathcal{H}_j$. The set of homographies associated to the target is $\{\mathcal{H}_t\}_t = \{\mathcal{H}_j, j = 1...J\}$.

For each frame, in its intermediate stages, the tracker predicts a BM using information from the previous output $\mathcal{M}(\Psi_{t-1}) : \mathcal{M}^P(\Psi_t)$.

5.3.3.2 System overview

Figure 5.5 depicts the three stages of the proposed method:

1. The **motion prediction** stage receives as inputs: the homographies defining the target motion in the previous frame, $\{\mathcal{H}\}_{t-1}$, and the output of the previous frame $\mathcal{M}(\Psi_{t-1})$. The output is a predicted mask $\mathcal{M}^P(\Psi_t)$ that establishes a potential location of the target in Ψ_t.

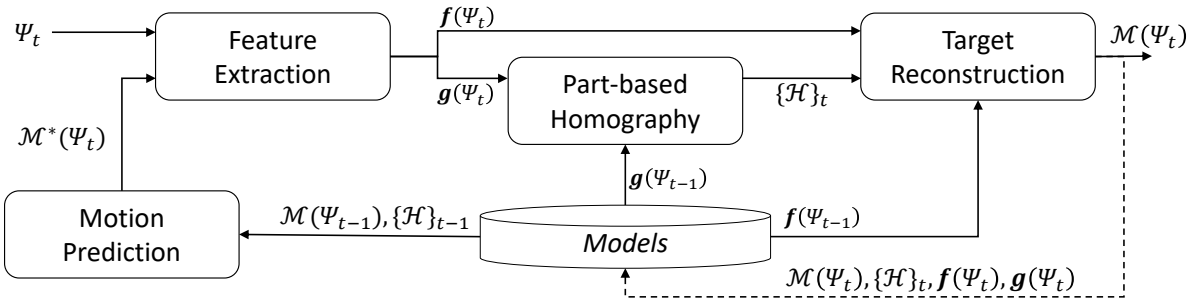


Fig. 5.5. System overview. It receives and input image (Ψ) and provides an BM output ($\mathcal{M}$).

2. The **feature extraction** stage takes as inputs: a new frame in the current time t, Ψ_t, and the predicted mask $\mathcal{M}^P(\Psi_t)$, output of the previous stage. It extracts cues, regions $\mathbf{f}(\Psi_t)$ and LF $\mathbf{g}(\Psi_t)$, and provides it as an output for subsequent stages.

3. The **part-based homography estimation** stage receives as inputs the region LF from LF extraction stage, $\mathbf{g}(\Psi_t)$, and the region LF of the previous frame, $\mathbf{g}(\Psi_{t-1})$. As output, it yields a set of homographies defining the target's parts motion respect to the previous frame, $\{\mathcal{H}\}_t$.

4. The **object reconstruction** stage receives as inputs the set of homographies $\{\mathcal{H}\}_t$, and the set of regions descriptions $\mathbf{f}(\Psi_t)$. As an output it provides a binary mask defining the shape of the target in the current frame, $\mathcal{M}(\Psi_t)$.

At the end of the process, the information provided to the next frame, lets call it models, is the information of the current frame, $\{\mathcal{M}(\Psi_t), \{\mathcal{H}\}_t, \{\mathbf{f}(\Psi_t)\}, \{\mathbf{g}(\Psi_t)\}\}$.

5.3.3.3 Motion prediction stage

The motion prediction stage aims to roughly estimate the target position in the current frame, $\mathcal{M}^P(\Psi_t)$.

For a certain part of the target, given $\mathcal{H}_j$and $\mathcal{M}(\Omega_{j,t-1})$ we use a Gaussian kernel, with standard deviation σ, to obtain $\mathcal{M}^P(\Omega_j)$ as a convolution $*$:

$$\mathcal{M}^P(\Omega_j) \coloneqq (\mathcal{H}_j x \mathcal{M}(\Omega_{j,t-1})) * G(\mathbf{x}, \sigma) \tag{5.2}$$

The predicted target position for Ψ_t is obtained as $\mathcal{M}^P(\Psi_t) = \bigcup_{j=1}^{J} \mathcal{M}^P(\Omega_j)$.

This process aligns with the idea of temporal coherence, i.e. target motion can be considered stable from frame to frame. However, to cope with sudden motion and scale changes, we smooth the previous location with the Gaussian prediction model BM. A example of motion prediction stage is shown in Figure 5.6.

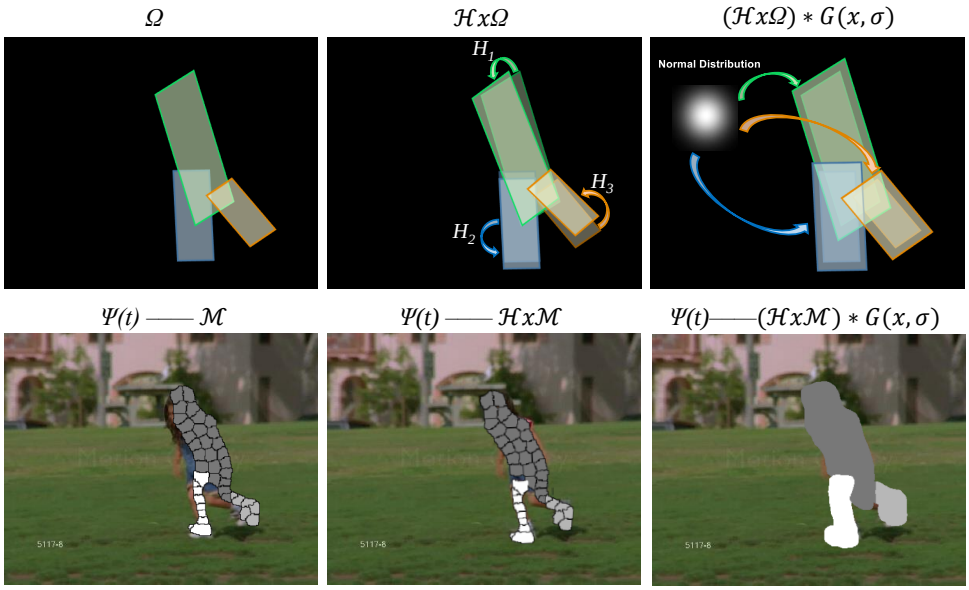


Fig. 5.6. Bottom row, overimposed on the input frame Ψ_t: part-wise target partition on the previous frame —a different gray level for each of the $J = 3$ parts. Composite regions of each part are indicated by black contours. Hypothesis/predicted mask obtained by using $\left\{\mathcal{H}_j^P, j = 1...J\right\}$ on each part. Final prediction location in the shape of a set of BM: $\left\{\mathcal{M}^P\left(\Omega_j\right), j = 1...3\right\}$.

5.3.3.4 Feature extraction

The feature extraction stage provides LF for: the homography estimation local—LF-DS $\mathbf{g}(\Psi_t)$—and for the target reconstruction—regions descriptions $\mathbf{f}(\Psi_t)$. It also defines the target parts spatial support in order to operate over them in subsequent stages.

Region extraction: It is based on Achanta [Achanta et al., 2012] SLIC superpixels technique. This technique relies on two parameters, the spatial-color ratio sp_r, and the size of the superpixels sp_n. Most of OTS superpixel-based techniques are defined as supervised because the need to fit those parameters depending the scenario. The proposal also includes those parameters, but the are unsupervisedly fixed.

- The sp_n defines the size of superpixels detected in the frame, i.e. the number of superpixels per frame. We aim to ensure that the target $\mathcal{T}$ is segmented in a minimum number of superpixels, $sp_n^{\mathcal{T}}$, so the further region based description is enough discriminative and repetitive. Being (w, h) the width and the height of the frame, $(w^{\mathcal{T}}, h^{\mathcal{T}})$ the same referred to the target, and the $sp_n^{\mathcal{T}}$ empirically fixed to 25, the sp_n parameter is defined as follows:

$$sp_n = sp_n^{\mathcal{T}} \frac{(w \cdot h)}{w^{\mathcal{T}} . h^{\mathcal{T}}} \tag{5.3}$$

- The sp_r defines the relation between spatial and color relevance for the segmentation process. Higher values means that spatial constrain is more relevant against color constrain, and vice versa. On one hand, if the spatial constrain is dominant, the algorithm will be better in segmentation challenges as camouflage, at the cost of loosing boundary accuracy. On the other hand, if the color constrain is more relevant, the fitting to the boundary will improve at the cost of being more vulnerable against segmentation challenges. The sp_r is fixed via a local maximization in the first frame. The maximized formula is the separability between foreground and background superpixels.

Local features extraction: The part-based homography relies on the SP-SIFT LF. The detection stage of this LF is controlled by a parameter defined as sp_{det}. To fix the value, a scan is performed using as stop condition the minimum number of detections required on each part of the target. The number is defined as 4, which is the minimum number of points required to define an affine homography.

Parts estimation: The target is divided into parts frame-by-frame. Those parts are theoretically defined as non-deformable areas of the target.

The part estimation process is defined as a superpixels clustering. Superpixels are clustered attending to color criteria, avoiding the inclusion of boundaries even if the color criteria suggest the merging. As a result of the process, the target is divided in a set of clusters. The criteria of not including boundaries in the clustering process guarantees non-deformable in short term, frame-by-frame, despite does not guarantee in long term. See Figure 5.7 for a simplified scheme of the part estimation process.

5.3.3.5 Part-based homography estimation

The part-based homography estimation stage aims to obtain a set of homographies, $\mathcal{H}^P$, defining the change, due to 3D motion, of the target parts.

Given $\mathbf{g}(\Psi_t)$ and $\mathbf{g}(\Psi_{t-1})$, a set of feature points description vectors extracted on Ψ_t and Ψ_{t-1}. Those LF are located only on the spatial support defined by the predicted location BM, and each feature point is associated to the part where it lies, $\mathcal{M}^P(\Omega_j)$.

The homography estimation is performed via matching of the feature points, which has proven to be a reliable technique to estimate homographies between pairs of images [Brown and Lowe, 2007], and it is faster than working at pixel level. However, two major problems appears when using feature points in tracking contexts:

- The blurring effect produced by motion is a major challenge for the detection of feature points. Detection techniques (e.g., DoG [Lowe, 2004], SURF detector [Bay et al., 2008],

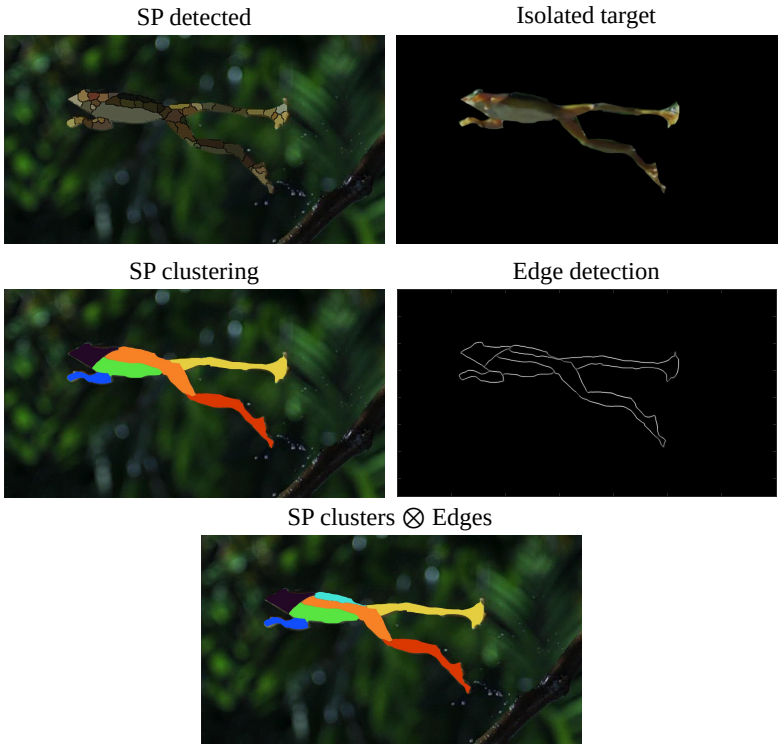


Fig. 5.7. Example part estimation. Left column SP detection and resulting clustering. Right column isolated target and resulting edges detection. Bottom row parts resulting of combining SP clusters and edges.

AGAST [Mair et al., 2010]) try to locate singularities on the gradient space of the image. Hence, as far as blurring attenuates gradients, the number of detections will decrease.

- The variability of the environment leads to a non discriminative description. As discussed in [Navarro et al., 2014b], the description certain of a local feature limiting a region belonging to background will include variable information as the target may be moving respect to the background.

To overcome these mayor challenges, a local optimization scheme is proposed to perform the part-homography estimation.

The scheme is divided into two steps: a direct homography estimation based on a random sample consensus scheme [Fischler and Bolles, 1981] and a back projection maximization to confirm the estimation (see Figure 5.8).

For a part j, the homography estimation step is based on a local feature matching among detected LF $\{\mathbf{g}(\Psi_t)\}$ and target model LF $\{\mathbf{g}(\Psi_{t-1})\}$. The matching strategy for the LF is $\mathcal{L}_2 - norm$, as proposed in the article [Navarro et al., 2014b]. Once the associations are defined, a random sample consensus algorithm is applied to diminish spatially the false positive matchings.

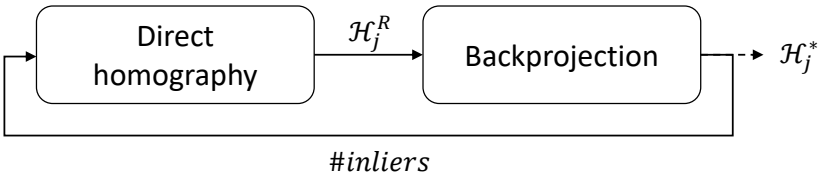


Fig. 5.8. Homography scheme for a certain part j. The random sample consensus stage proposes homographies and the Back-projection obtains an agreement measure ($\#inliers$). The process iterates until the agreement measure finds a maximum value.

The spatial translation defined with the LF matchings is used to propose an homography as candidate, $\mathcal{H}_j^R$. The proposed homography defines an affine transformation between target parts location in the previous frame and in the current frame.

The back projection stage inverts the homography matrix and maps target parts (its pixels) in the current frame to positions in the previous frame. A pixel level distance is calculated between the pixels in the previous frame and the back-projected pixels. Each position is considered an inlier if its distance is under a maximum value.

The two steps process iterates finding a maximum in the number of inliers. This process is processed in parallelized for each part of the target.

As an output of the part-based homography stage, a set of homographies $\mathcal{H}^P$ is defined.

5.3.3.6 Target reconstruction

The target reconstruction stage has two purposes.

1. Refine the hypothesized location, $\mathcal{M}^P(\Psi_t)$, to provide tight-to-object tracking results.

2. Obtaining a reliable segmentation onto which update the target and background models for the next frame.

Let $\left\{\mathcal{H}_j^P, j = 1....J\right\}$, be the set of estimated homographies, and let $\left\{\mathbf{f}(r_{\Psi_t,k}^{\mathcal{T}})\right\}$ and $\left\{\mathbf{f}(r_{t,k}^{\mathcal{B}})\right\}$, be the set of region's descriptions vectors of the background and target respectively.

This last obtained as described in Section 5.3.3.1, and using the hypothesized location $\mathcal{M}^P(\Psi_t)$ to define the set of regions in which the target is partitioned in the current frame.

The target regions are then back-projected to the previous frame, using the transposed set of homographies $\left\{\mathcal{H}_j'^P, j = 1....J\right\}$.

Then, these back-projected regions are assigned, under a subjective assignation and a nearest-neighbor premise, either to the set of background or target regions.

Finally, the set of regions classified as target are identified in the current frame and used to

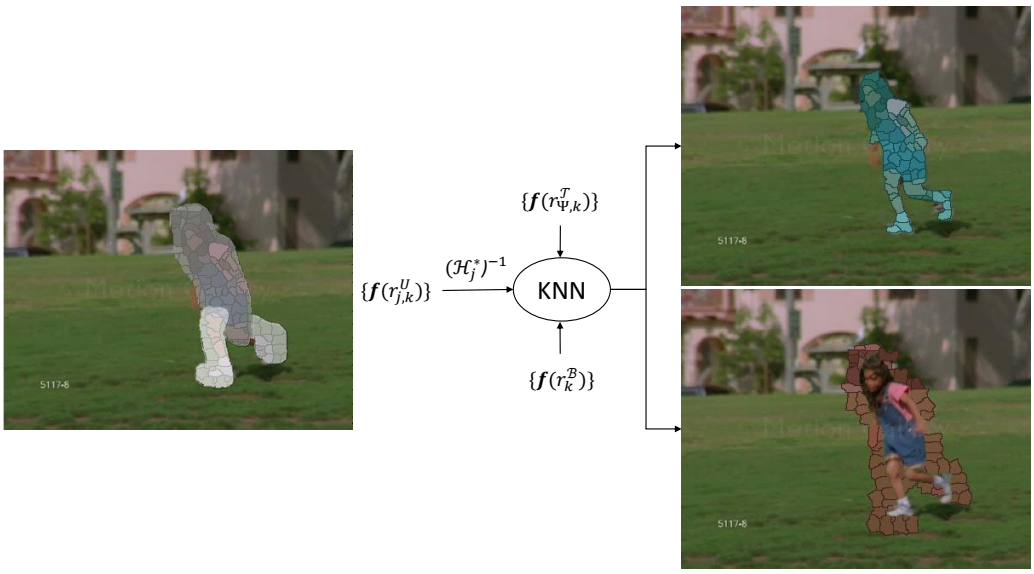


Fig. 5.9. Object reconstruction scheme. Regions are detected in the current frame, and associated to a part depending on their spatial position. Regions are classified in foreground or background using a KNN classifier. Results of the classification are shown in the last column, blue defines the regions classified as target, and pink defines the regions classified as background.

define the output mask:

$$\mathcal{M}(\Psi) = \bigcup_{j=1}^{J} \left\{ \mathbf{f}(r^{\mathcal{T}}_{\Omega_j,k}), k = 1...K_j \right\} \tag{5.4}$$

The process is illustrated in Figure 5.9.

5.3.4 Experimental evaluation

5.3.4.1 Evaluation purposes

The proposal can be defined as an OTS-like approach. However, as claimed in section 5.3.1, it aims to cope also with classical tracking challenges, rather than focusing only on segmentation challenges. To fulfill these premise, and to generally evaluated the designed approach we define two major goals:

- The first one is to provide comparable tracking accuracy in segmentation tasks to OTS algorithms,i.e. a tracking by segmentation evaluation, focusing on the obtention of Binary Masks (BM).

- The second is to show that the proposal improves both OTS and BB tracking algorithms under classical tracking challenges following a tracking by bounding box (BB) protocol.

To fulfill both goals, two evaluation scenarios are defined. On them, different datasets, algorithms, metrics and results are defined according to the evaluation objectives: tracking by segmentation (BM) evaluation and tracking by bounding box (BB) evaluation.

5.3.4.2 Tracking by segmentation evaluation

Dataset Tracking by segmentation datasets are complex to generate. They are usually focused on segmentation challenges as camouflage or clutter. They are annotated at pixel level and usually frame by frame. Due to the high cost of generating these datasets, there are few available sets in the state of the art. Since the emergence of SegTrack_v2 [Li et al., 2013], most of the recent tracking segmentation approaches have used it as the reference dataset to compare with the state-of-the-art. To fairly compare our proposal against state-of-the-art, the same sequences analyzed in [Li et al., 2013] are used in the evaluation. We name this set of videos and associated annotations the SegTrack_v2 dataset.

The benchmark is composed of 14 sequences, up to a total of 976 manually annotated frames. Thumbnails of the dataset are shown in the Figure 5.10. Tags below each thumbnail are used to identify each sequence in subsequent sections. In the sequences with multiple annotated objects, results are disaggregated for each one, summing up to a total of 24 objects to be tracked.

In [Li et al., 2013], authors propose to evaluate trackers under challenges as motion blur, appearance change, complex deformation, occlusion, slow motion and interacting objects. However, in the light of the sequences, the main challenges in this dataset are camouflage and clutter, as the other listed tracking challenges are of low-complexity compared to those in Bounding Box oriented datasets.

Approaches evaluated for comparison Whereas, there are several surveys [Smeulders et al., 2014] and contests [LIRIS, 2014] which help to assemble state of the art in BB-based algorithms, there is a lack of comparative evaluations and contests in the OTS field, leading to a weakly categorized state-of-the-art. To cope with this issue, we rely on the algorithm selection made in [Wen et al., 2015]. Table 5.3 lists these algorithms and also includes a brief summary of their main characteristics in terms of supervision, online behavior, segmentation technique and representation of the segmentation process. Algorithms defined as unsupervised are those who do not need parameter tunning depending on the scenario or sequence. Algorithms defined as online are those that provide results frame by frame, and not at the end of the whole sequence. As shown in the summary, there are four unsupervised algorithms and three supervised. Among the unsupervised approaches there is also an offline algorithm. To out knowledge, these seven algorithms encompass the main strategies of the related work in tracking by segmentation.

[1] Composite statistical interference

Fig. 5.10. SegTrack_v2 dataset. Each thumbnail correspond to a sequence of the dataset. The labels will identify each sequence in the subsequent evaluation.

Experimental setup We propose to evaluate BM-based approaches under the classical Intersection over Union measure, $\cap/\cup$. In order to globally asses the operation of a given an algorithm, we propose to characterize the algorithm by the variation coefficient $c_{\cap/\cup}$(see equation 5.5).

$$c_{\cap/\cup} = 100(1 - \frac{\sigma_{\cap/\cup}}{\mu_{\cap/\cup}}) \tag{5.5}$$

, where $\sigma_{\cap/\cup}$ and $\mu_{\cap/\cup}$ are the standard deviation and average of the $\cap/\cup$ results obtained for a the algorithm along all the sequences in the the SegTrack_v2 dataset. The use of the variation coefficient, instead of simply averaging its $\cap/\cup$ results, aims to fairly condense an algorithm operation in a single measure, avoiding the overrating of polarized algorithms yielding excellent results in some sequences and poor results in some others.

Both measures, $\cap/\cup$ and c_v are defined as in [Li et al., 2013]. For the seven algorithms,

Algorithms	Ref	Unsupervised	Online	Segmentation	Representation
KeySeg	[Lee et al., 2011]	✓	✗	Region ranking	BPLR
EGraph	[Grundmann et al., 2010]	✓	✗	spatial-temporal graph	Spatial + Optical flow
SPT	[Li et al., 2013]	✓	✗	Figure-ground	Color SIFT
SPT+CSI	[Li et al., 2013]	✓	✓	Figure-ground + CSI[1]	Color SIFT
HB	[Godec et al., 2013]	✗	✓	Grab-cut	Hough forest
DynGraph	[Cai et al., 2014]	✗	✓	Sp + DinamicGraph	Spectral centroid
JOTS	[Wen et al., 2015]	✗	✓	SLIC Sp	HSV + Spatial
HPSTr	Proposal	✓	✓	SLIC Sp	RGB + SP-SIFT

Table 5.3: OTS algorithms selected for evaluation. Main stages description. CSI means Composite Statistical Interference. Sp means superpixels. BPLR means Boundary Preserving Local Regions.

figures in Table 5.4 are from their original papers or from the comparison made in [Li et al., 2013].

Discussion **Overall results:** Attending to the overall results per category, our proposal is the best among unsupervised algorithms, close to JOTS which is the top approach among supervised algorithms. SPT and SPT+CSI also achieve highly ranked results. The proposed approach operates better than the other unsupervised and online algorithm (SPT) in 8 out of 14 sequences, and generally improves its average operation.

Our approach and JOTS rank top, each, in four out of 14 sequences. SPT+CSI and KeySeg get the best results in three sequences each one.

Scale and shape: Although these challenges are present in almost every sequence, they are specially remarkable in *worm, bird of paradise, monkeydog* and *hummingbird* sequences. In *worm* and *bird of paradise* sequences, objects are subjected to sudden scale changes and complex shape deformations under little target displacement. In the *monkeydog* and the *hummingbird*, besides the scale and deformation changes, the target is also subjected to sudden displacements. These sequences require a trade-off between segmentation robustness to hold shape deformations, and tracking robustness to hold scale changes and sudden motion. Our approach, ranks first in these sequences, as is able to balance both tasks effectively thanks to the region and local feature approach.

Camouflage: Most challenging sequences in camouflage aspect are the *cheetah*, *parachute*, *birdfall* and *monkey* sequences. In these sequences, the tracker operation is strongly conditioned by the segmentation results. The supervised algorithm JOTS, with parameter specially tuned to these sequences provides better figures than the rest of the algorithms. The overall operation of the proposed approach is biased by the low figures obtained for the *birdfall* sequence, where the movement of the target in the scene is almost nonexistent, the target is so small that the LF are severely affected, and the clutter an camouflage challenges are hard.

Sequences\Algorithms	**HB**	**DynGraph**	**JOTS**	**KeySeg**	**EGraph**	**SPT+CSI**	**SPT**	**Proposed**
Unsupervised				✓	✓	✓	✓	✓
Online	✓	✓	✓				✓	✓
Worm	36,8	44,3	79,3	**84,4**	34,7	82,8	75,6	72,3
Bird of paradise	5,1	46,5	93,0	92,2	86,8	94,0	88,2	**95,1**
Frog	14,5	38,8	56,3	0,0	67,1	72,3	65,8	**74,0**
Monkeydog*	40,4	14,8	51,6	39,6	43,5	45,1	38,1	**68,0**
Hummingbird*	5,9	48,1	54,5	60,1	19,4	63,3	55,3	**63,5**
Cheetah-deer*	46,7	29,2	**50,7**	28,1	21,6	39,1	40,7	41,2
Parachute	85,6	59,3	94,4	**96,3**	69,1	93,4	93,2	94,6
Penguin*	35,8	30,8	**88,9**	9,3	74,4	63,0	60,6	86,2
Birdfall	56,0	36,4	**78,7**	49,0	57,4	62,5	62,0	56,4
Monkey	73,1	58,7	**86,0**	79,0	61,9	84,8	84,1	71,4
Girl	53,6	62,0	84,6	87,7	31,9	**89,2**	89,1	74,9
Soldier	70,7	54,2	81,1	66,6	66,5	**83,8**	83,0	56,6
Drift-car*	42,2	35,5	65,5	46,9	41,2	**67,7**	66,1	48,2
Bmx*	1,0	19,9	47,3	**63,0**	35,8	55,1	49,8	50,5
$c_{\cap/\cup}$	*34,4*	*64,6*	***76,2***	*47,1*	58,8	*75,4*	*73,4*	***75,6***

Table 5.4: BM evaluation results. Intersection-over-union metric in the SegTrack v2 dataset. Values in bold shows best results per sequence. Bottom line shows the overall operation results, in bold the best supervised results and the best unsupervised. * Average results of each target tracked in the sequence.

Clutter: The other sequences: *girl*, *drift-car*, *soldier* and *bmx* present challenging clutter situations produced by an heterogeneous and deformable target which inhibits the learning of a representative model. SPT-related approaches present the best behavior in these situations, specially in the offline version (SPT+CSI). The KeySeg approach also performs well in those situations. Apparently, the clutter challenge is better faced by an offline strategy, as in this case, the target model can be learned in consonance with the whole set of target appearances along the sequence. In contrast, the proposed approach evidences some problems in these situations. For instance, the *soldier* sequence includes camouflage and sharped contours defining a big target, hence the unsupervised parameter setting fails to define small regions for sharped contours in such a big target. The *drifting car* and the *bmx* sequence present two targets moving in the same direction and speed. The approach tends to merge them considering the second target as unseen regions of the first one.

In overall, our approach has been successfully tested against state of the art datasets and algorithms. Obtained results lead to the conclusion that HPSTr segmentation results are as

Fig. 5.11. BB-challenges dataset. Each thumbnail correspond to a sequence of the dataset. The labels will identify each sequence in the subsequent evaluation.

good or comparable to top supervised and/or offline approaches in state of the art, being our proposal unsupervised and online. However, it is fair to conclude, that if online and unsupervised operation is not an issue, JOTS operates slightly better than the proposed approach albeit JOTS is less scalable and generic because it benefits from a tailored setup for each video.

5.3.4.3 Tracking by bounding box evaluation

Dataset There is a large number of proposed datasets for evaluating BB-based tracking algorithms, including PETS [Patino et al., 2016], VOT [Kristan et al., 2015] and [Yang et al., 2014]. In [Smeulders et al., 2014], the common challenges that should be handled by BB-based tracking algorithms are exhaustively described. Besides, we also consider that segmentation-related challenges as camouflage and shape deformation need to be also accounted. With this in mind, we choose the dataset proposed in [Yang et al., 2014] as a suitable set for evaluation as it provides a well balanced set between diversity and complexity of tracking challenges, including common BB and segmentation-related challenges. Specifically, the 10 sequences in the dataset cover the following tracking challenges: complex background, moving camera, fast target movement, large variation in target's pose and scale, half or full occlusion of the target, target deformation and distortion. See Figure 5.11 for example frames of each sequence.

Approaches evaluated for comparison We propose to compare our method against state-of-the-art matching and discriminative based trackers. Among the matching based trackers we include classical algorithms like Mean Shift (MS) [Collins, 2003] and Particle Filter (PF) [Nummiaro et al., 2003], as well as OLT-matching based algorithms as IVT [Lim et al., 2004], Frag [Adam et al., 2006], PROST [Santner et al., 2010] and VTD [Kwon and Lee, 2010]. On the discriminative based category, we evaluate the behavior of foreground-learning trackers as:

MIL [Babenko et al., 2009], TLD [Kalal et al., 2010] and Struck [Hare et al., 2011], as well as that of foreground and background-learning trackers including SPTrack [Wang et al., 2011] and R-SPTrack [Yang et al., 2014]. The proposed approach lies in this last category. Furthermore, due to its success in the tracking-by-segmentation experiment (see section 5.3.4.2 and Table 5.4), we also evaluate the operation of JOTS [Wen et al., 2015] in this dataset.

Experimental setup The comparison is performed in terms of the PASCAL VOC detection criteria [Everingham et al., 2010]. Given a tracker and a sequence, the detection criteria (stf) is defined as the total number of frames of the sequence where the tracker's output satisfies the PASCAL overlapping criteria. To faithfully asses the tracker behavior with independence of the sequence length, the detection criteria is usually expressed in percentage terms, $stf_{\%}$. When available, tracking results are extracted from the original paper or from the comparison made in [Yang et al., 2014]. For JOTS and other supervised approaches, the algorithm is run on the dataset using the default parameters suggested by the authors. Additionally, in order to globally asses the operation of a given algorithm, we propose to characterize it by the variation coefficient c_{stf} (see equation 5.6).

$$c_{stf} = 100(1 - \frac{\sigma_{stf}}{\mu_{stf}}) \tag{5.6}$$

, where σ_{stf} and μ_{stf} are the standard deviation and average of the stf results obtained for a the algorithm along all the sequences in the dataset. Results are included for comparison in Table 5.5 and Table 5.6. In the first one, the proposal is compared with the matching-based trackers. In the second, against the discriminative-based trackers and the OTS JOTS.

Discussion **Overall results:** Attending to the variation coefficient, our proposal ranks best among the evaluated algorithms in successfully tracked frames. Besides, it leads the comparison in six out of 12 sequences. Each one of the SPTrack and its evolution, the R-SPTrack, leads two other sequences. Finally TLD ranks best in the other two sequences. Comparing the tracking strategies, results are in consonance with the ideas in [Smeulders et al., 2014]: discriminative trackers outperform matching based approaches. Among the pool of matching algorithms, the OLT approach Frag is the one with best results. Among discriminative approaches, results of the target and background based algorithms are far better than the obtained by the target model based ones. Finally, the unsupervised version of the JOTS tracker yields low figures, suggesting that the algorithm is highly sequence-dependent and that it is able to effectively cope with segmentation challenges but not with classical tracking challenges.

Occlusions: This is one of the main classical tracking challenges. In the dataset, occlusions are present in sequences: *woman, liquor, lemming, girl, bird1, bird2* and *basketball*. Our proposal presents the best results in average in these sequences. The robust to occlusion LF used in the

Algorithms	**MS**	**PF**	**IVT**	**Frag**	**PROST**	**VTD**	**Proposal**
Nature	Matching						Discriminative
Learning	None		Target				Target & Bkg
woman	35 (7,95)	31 (7,05)	49 (11,14)	44 (10,00)	-	27 (6,14)	313 (71,14)
liquor	413 (23,72)	1.202 (69,04)	380 (21,83)	1.375 (78,98)	1.444 (82,94)	471 (27,05)	**1.719 (98,74)**
racecar	43 (5,78)	207 (27,82)	17 (2,28)	111 (14,92)	-	42 (5,65)	**519 (69,76)**
lemming	171 (12,80)	426 (31,89)	1.046 (78,29)	678 (50,75)	969 (72,53)	471 (35,25)	**1.295 (96,93)**
girl	79 (5,27)	1.106 (73,73)	107 (7,13)	628 (41,87)	-	828 (55,20)	1.192 (79,47)
singer1	64 (18,23)	96 (27,35)	332 (94,59)	87 (24,79)	-	350 (99,72)	316 (90,03)
bird1	1 (0,25)	6 (1,47)	4 (0,98)	47 (11,52)	-	7 (1,72)	134 (32,84)
bird2	36 (34,95)	19 (18,45)	9 (8,74)	42 (40,78)	-	9 (8,74)	**97 (94,17)**
basketball	78 (10,76)	455 (62,76)	80 (11,03)	512 (70,62)	-	601 (82,90)	**715 (98,62)**
bolt	15 (4,29)	172 (49,14)	4 (1,14)	32 (9,14)	-	195 (55,71)	241 (68,86)
transformer	28 (22,58)	32 (25,81)	29 (23,39)	38 (30,65)	-	47 (37,90)	120 (96,77)
surfing1	36 (12,50)	16 (5,56)	24 (8,33)	28 (9,72)	-	24 (8,33)	100 (34,72)
c_{stf}	*24,68*	*25,16*	*0*	***26,00***	*0*	*8,19*	***69,72***

Table 5.5: BB evaluation results. Matching trackers are compared with the proposal. Results are shown as: successfully tracked frames (percentage successfully tracked frames), metrics in the [Yang et al., 2014] dataset. Overall results in variation coefficient terms.

Algorithms	**MIL**	**TLD**	**Struck**	**SPTrack**	**R-SPTrack**	**Proposal**	**JOTS**
Nature	Discriminative						OTS
Learning	Target			Target and Background			
woman	38 (8,64)	36 (8,18)	**333 (75,68)**	310 (70,45)	298 (67,73)	313 (71,14)	-
liquor	353 (20,28)	1.398 (80,30)	405 (23,26)	1.701 (97,70)	1.698 (97,53)	**1.719 (98,74)**	544 (31,25)
racecar	33 (4,44)	24 (3,23)	51 (6,85)	340 (45,70)	345 (46,37)	**519 (69,76)**	56 (7,53)
lemming	1.105 (82,71)	361 (27,02)	652 (48,80)	1.290 (96,56)	1.277 (95,58)	**1.295 (96,93)**	-
girl	560 (37,33)	169 (11,27)	246 (16,40)	1.180 (78,67)	**1.439 (95,93)**	1.192 (79,47)	-
singer1	84 (23,93)	**351 (100)**	87 (24,79)	297 (84,62)	347 (98,86)	316 (90,03)	222 (63,25)
bird1	114 (27,94)	25 (6,13)	17 (4,17)	**139 (34,07)**	84 (20,59)	134 (32,84)	33 (8,10)
bird2	86 (83,50)	12 (11,65)	14 (13,59)	94 (91,26)	90 (87,38)	**97 (94,17)**	17 (16,50)
basketball	204 (28,14)	46 (6,34)	85 (11,72)	707 (97,52)	695 (95,86)	**715 (98,62)**	-
bolt	12 (3,43)	49 (14,00)	9 (2,57)	224 (64,00)	231 (66,00)	241 (68,86)	**301 (86,00)**
transformer	30 (24,19)	43 (34,68)	34 (27,42)	**124 (100)**	**124 (100)**	120 (96,77)	-
surfing1	10 (3,47)	**116 (40,28)**	24 (8,33)	80 (27,78)	98 (34,03)	100 (34,72)	29 (10,07)
c_{stf}	*4,85*	*0*	*3,35*	*65,02*	*62,65*	***69,72***	*3,40*

Table 5.6: BB evaluation results. Discriminative trackers and the OTS JOTS are compared with the proposal. Successfully tracked frames (percentage successfully tracked frames) metrics in the [Yang et al., 2014] dataset. Overall results in variation coefficient terms.

part-based homography stage, and the discriminative superpixels based strategy are the main reasons. This second statement is also supported in the results of SPTrack and R-SPTrack. PROST algorithm present good results in an occlusions sequence. It has a non-adaptive template matching stage, very robust when the target undergoes no major appearance change, as it happens in the liquor sequence. In tracking strategies terms, matching a target which may be partially occluded is a hard task. Thus, the discriminative trackers generally yield better results in this challenge.

Clutter and camouflage: These challenges are mainly represented in sequences *lemming*, *girl* and *singer1*. In these sequences, the clutter and camouflage is combined with motion or scale changes. The R-SPTrack outperforms the other algorithms as it is mainly based in a robust to camouflage segmentation approach. Our proposal fails in sequences like *singer1*, as it presents an homogeneous target where LF-DS stage present difficulties. In overall, the presence of high complex scenarios and non discriminative targets tend to worsen discriminative approaches results in comparison with matching based approaches.

Sudden motion: Trackers presenting motion prediction stages struggle in the presence of sudden motion situations. The sequences *bird1*, *bird2* and *basketball* are examples of this problem, where the proposed approach yields good results in comparison with the other evaluated trackers. This can be explained by the object reconstruction stage of the proposal, which apparently is able to correct inadequate target displacement estimations made in the motion prediction stage. The simple Gaussian motion estimation followed by SPTrack appears to also effectively handle the sudden motion challenge. Results in this category are similar for both matching and discriminative strategies, and differences may only rely on the presence of other challenges in the sequence as camouflage or occlusions.

Appearance changes: Learning approaches aims to adapt their model to be robust to this target appearance changes. However, if the change is fast (as in the *transformer* sequence) or if the target not only changes in appearance but also is affected by sudden motion or occlusions (as in the *surfing1* sequence), appearance changes are highly problematic even for OLT approaches. The R-SPTrack, SPTrack and our approach are the only able to handle the high complex appearance change suffered by the target in the sequence transformer. Regarding the proposed approach, the matching of LF is able to handle the sudden motion between frames despite the structural deformation of the target. On the other hand, no algorithm is able to reach a 50% of successfully tracked frames in the *surfing1* sequence. This leads to the conclusion that appearance change is a major challenge for any evaluated tracking algorithm. This may be a consequence of the tug-of-war between discrimination and adaptability: increasing the algorithm discriminativeness to be robust to most of the tracking challenges usually leads to a higher sensitivity of the method to appearance changes, and viceversa.

5.3.5 Discussion of the SP-SIFT as the core of the application experiment

In conclusion, the proposed approach outperforms evaluated state of the art BB-trackers in classical tracking challenges. It also outperforms the other OTS approach evaluated in this section, suggesting that the majority of OTS approaches are focused on handling segmentation-related challenges rather than tracking-related. Apart from our proposal, SPTrack and R-SPTrack present also remarkable results in the evaluation. They share with the proposal the discriminative strategy and the SLIC superpixels segmentation stage. Therefore, discriminative strategies using regions—SP-SIFT—have demonstrated to be reliable techniques to face classical tracking challenges.

Part IV

Enhancing the scale adaptation of region segmentation algorithms via local features

Chapter 6

The LF-SLIC algorithm: Enhancing the discriminative capacity of a region segmentation via local features

In this chapter we present the proposed method LF-SLIC. The method aims to validate one of the thesis statements: region segmentation algorithms can be enhanced by combining them with local features. First, we study the few reference methods in the literature following similar ideas. Then, we present the proposed method and describe the resulting solution. Finally, we present the concept validation test. This chapter describes the work that led to the international journal article : "Accurate segmentation and registration of skin lesion images to evaluate lesion change", [Navarro et al., 2018a].

6.1 Method motivation

The objective of the proposed method is to validate the second of the statements presented in the thesis hypothesis. The statement suggest that the capabilities of LF and region segmentation algorithms are complementary. Specifically, it claims that LF can improve region segmentation algorithms via increasing their adaptation to the scale information.

As with the SP-SIFT feature, the idea is not only to propose a specific method to validate the statement. It is also to establish a standard to combine region segmentation algorithms and LF.

Superpixels have proven to be a top region segmentation technique in terms of boundary adherence. Among the existing algorithms, SLIC superpixels [Achanta et al., 2012] can be

considered the reference in terms of performance, speed and contrasted evaluations. As discussed in Chapter 2, region segmentation algorithms using spatial constraints obtain better overall results. However, smallest or biggest areas of the image can be under or over segmented. This is caused by a poor scale management, which is common in these techniques. LF are one of the most effective approaches in the state of the art for managing the scale-space concept, especially in the detection stage.

It is accepted that on almost every image there will be objects of different sizes. In region segmentation, using the same spatial constraint for the whole image leads to a poor segmentation in the boundaries of the smallest objects, and to an oversegmentation in the biggest ones, typically large homogeneous areas. We discard the idea of oversegmenting the image to cope with the smallest object. Oversegmenting will result in increasing the computational cost to the limit of working with regions of pixel size. The proposal is to apply different spatial constraints—for one image—in the segmentation process. Each constraint will be decided using the scale information linked to the LF detections. We can define an ideal segmentation of an image as the one able to get the best boundary adherence with the minimum number or regions.

This situation is illustrated in Figure 6.1. An image is segmented using SLIC superpixels method. We use different spatial constraints. The smallest regions are responsible for the details of the dog's face. The medium size regions accurately define the dog's body. The biggest regions are the best ones for segmenting the non-relevant-information background. Thus, an *ideal* segmentation might combine the spatial constraints of the algorithm, resulting in an accurate segmentation where required and saving resources where possible. Using the proposed method, detailed below, the objective is to get this *ideal* segmentation. We will evaluate the quality of the segmentation in terms of the properties presented in Chapter 2.

The proposed LF-SLIC method (which respond to Local Feature based SLIC), is defined using the superpixels segmentation algorithm SLIC [Achanta et al., 2012], and the SIFT feature detector [Lowe, 2004]. However, the concept of using LF in the region segmentation scale management can be done using different combinations of segmentation algorithms and LF. Our results indicate that this technique achieves:

- The highest boundary adherence that the SLIC superpixels can get in the target details.

- A remarkable computational cost saving, and a notable reduction of application distractors in homogeneous and/or out of the target areas.

Fig. 6.1. *Ideal* segmentation example. Left column, top to bottom: original image, and image segmentations using large, medium and small spatial constraints. Right column, top to bottom: ideal segmentation resulting from the combination of segmentations per constraint, and the part of each segmentation, with different spatial constraint, that contributes to the ideal segmentation.

6.2 Background on region segmentation algorithms enhanced with local features

We need to define first our understanding of multiscale segmentation in the context of this thesis. Not all the region segmentation algorithms can be considered multiscale. In fact, only those with some kind of spatial constraint. There are a lot of techniques whose segmentation criteria are not spatial. These techniques can provide regions of different sizes, but there is no multiscale process, just the result of the merging criteria.

To our knowledge, there is no spatially-constrained constrained region segmentation approach using LF as we propose or similarly. There are several solutions proposing to use LF on top of superpixels, but no reference proposes the reverse approach. We then include solutions proposing multiscale superpixel segmentation, which are somehow close to the idea behind our proposal. Most of these approaches are working with video segmentation, where the computational cost associated with oversegmentation is critical. The most relevant or related approaches are described here:

1. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images [Bejnordi et al., 2015]. It obtains excellent segmentation results, but the process is semi-supervised. Superpixels are scaled as the user zooms in certain areas of the image.

2. Efficient Hierarchical Graph-Based Video Segmentation (EGraph) [Grundmann et al., 2010] is one of the most referenced superpixels segmentation technique and one of the first using multiscale techniques. They propose a graph-based approach to merge regions with inconsistent edges between scales, texture information and a maximum size threshold. Their results in proper segmentation evaluations are behind techniques in the state-of-the-art as SLIC, but its mayor drawback is the computational cost.

3. Multiscale Symmetric Part Detection and Grouping (MSPDG) [Levinshtein et al., 2013] proposes a multiscale superpixel segmentation. The scales merging is graph-based for skeletonization. Their results are state of the art, but their applications are restricted to people related tasks.

4. Edge-Weighted Centroidal Voronoi Tessellations (EWCVT) [Wang and Wang, 2012] and its evolution Hierarchical Edge-Weighted Centroidal Voronoi Tessellations (HEWCVT) [Zhou et al., 2015]. They propose a multiscale superpixel segmentation based on the concept of Centroidal Voronoi Tessellations. The technique guides a merging of the superpixels initially generated at the smallest size into bigger regions. Their main objective is to reduce the computational costs, but not to improve the segmentation results for targets at different scales in the image.

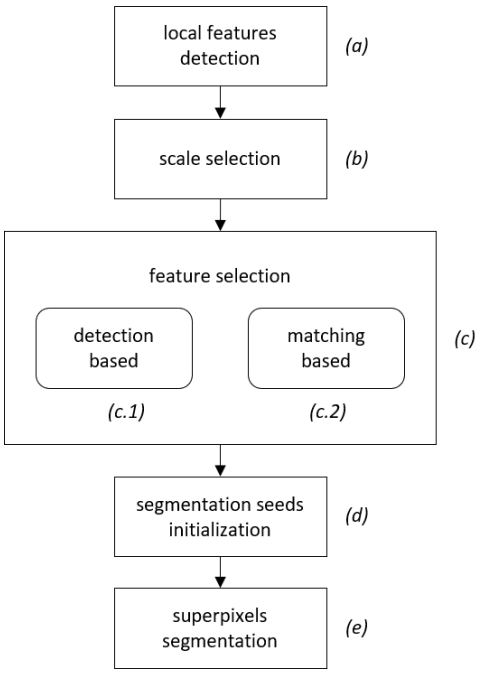


Fig. 6.2. LF-superpixels segmentation: method overview. The process consists of five main stages: (a) local feature detection; (b) scales selection based on the LF detected; (c) feature selection, which depending on the application will be: (c.1) detection based, or (c.2) matching based; (d) segmentation seeds initialization using the spatial location of the selected LF; (e) superpixel segmentation, based on seeds initialization.

The conclusion of the state-of-the-art review is the lack of similar approaches, and the multiscale segmentation as a way of reducing computation costs in the corresponding applications.

6.3 The LF-SLIC method

Figure 6.2 shows a diagram of the method. We will use the indexes in the Figure to guide the method description. Notice that, as in Chapter 4, the schema is defined for a generic region segmentation algorithm, with the restriction of being based on seeds initialization and spatial constraints. In particular, the method described below describes the proposed schema for the SLIC region segmentation algorithm and the SIFT feature. Figure 6.3 will be additionally used to graphically support the description process and to compare the two segmentation results, the one obtained with the original SLIC technique versus the obtained with the proposed LF-SLIC.

6.3.1 LF-SLIC proposed schema

The original SLIC technique has been shown to be highly competitive for image segmentation. However, if there is a region of interest (ROI) in the image, segmenting the whole image is

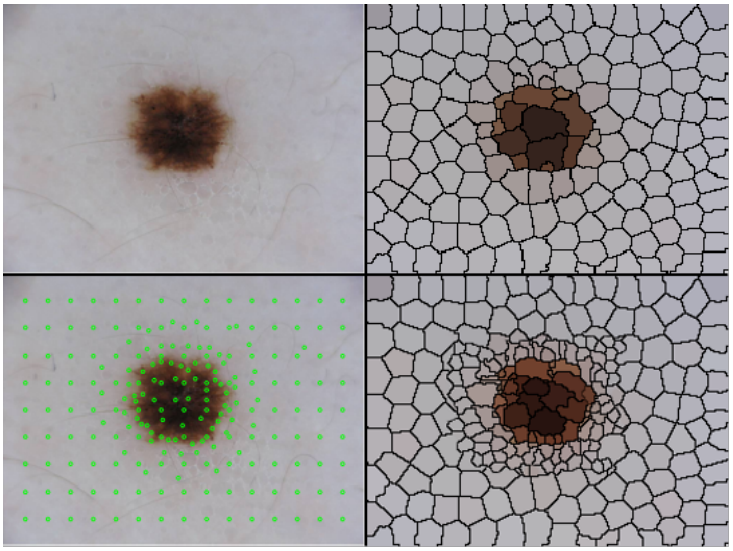


Fig. 6.3. SLIC versus LF-SLIC visual comparison. Top-left input image. Top-right SLIC segmentation result. Bottom-left LF-SLIC initialization seeds. Bottom-right LF-SLIC segmentation result.

useless and accurately defining the contour of the ROI is convenient. SLIC original methods propose to initialize the region centers or seeds, using a regular grid (see results in Figure 6.3 top right image). We can appreciate that the SLIC segmentation is missing small details in the skin lesion boundaries whereas it extracts useless boundaries in the rest of the image. We propose to replace uniform seeds initialization with feature-driven initialization, so that superpixels are forced to be smaller around detected LF (see results in Figure 6.3 bottom row). To do so, we propose a schema that slightly varies which detections the segmentation algorithm uses: a detection-based initialization, or a matched detection-based initialization. The decision depends on the application. Following the naming in Figure 6.2 the process consists in:

(a) Local feature detection. Given an input image, we process it to obtain the local feature detections. According to the proposed technique in SIFT, the resulting detections will be pairs location-scale $((x, y), \sigma)$. The scale value indicates at which scale the feature is detected: the higher the scale, the smoother the image (which SIFT manages as a smaller image). SIFT detection process defines for every image the following scales to be analyzed: σ_0 - corresponding to an image twice bigger than the original; σ_1 - corresponding to the original image; and σ_{02-N} where $N = 4$ - each image corresponding to subsampled versions of the original. The output of this stage is a number of feature locations with an associated scale value.

(b) Scale selection. We compare all the detections obtained in the previous stage. We eliminate those that overlap with another detection in a lower scale. The overlapping criteria is based

on the description area of the local feature. LF whose description area overlaps more than 50% with the description area of a feature in a lower scale are discarded.

(c) LF selection. As mentioned before, depending on the application we have two different processes. If we are segmenting an image without further information about which is the target, we follow (c.1). If we are segmenting an image with a known target, i.e. we have previous feature descriptions of the target, we follow (c.2).

(c.1) Detection based. All the detected LF are selected for the next stage.

(c.2) Matching based. We match the detected LF with the previously obtained target LF. Positive matchings are the LF selected for the next stage. For the matching process we follow the proposed technique in the original SIFT paper [Lowe, 1999].

(d) Segmentation seeds initialization. According to the SIFT operation, in the set of selected LF we have different scales: $\{\sigma_0 = 2; \sigma_1 = 1; \sigma_2 = 0.5; \sigma_3 = 0.25; \sigma_4 = 0.125\}$, where the value is the scaling factor. We first create a set of segmentation seeds initialization, one per scale. Instead of resizing the image, we vary the number of regions to be obtained (SLIC parameter): we increase the number for lower scales (biggest images) and reduce it for higher scales (smaller images). The results are shown in Figure 6.4. Then LF are used to define which seeds we use on each area of the image. We have two possibilities here: 1) Pixels in the image not associated to any detected feature: we use the seeds of the higher scale. 2) Pixels in the image associated to a feature: we use the seeds of the scale of the feature. The results are an initialization grid like the one shown in Figure 6.3.

(e) SLIC superpixels segmentation. We follow the original paper segmentation process starting right after the seeds initialization. Further details can be found in the original paper [Achanta et al., 2012].

6.4 Proposal validation test

We define two aspects that need to be tested in order to validate the proposal. Application-level validation will be performed in Chapter 7. The first aspect is to test whether the proposal maintains or improves the segmentation capabilities, i.e. the boundary adherence of the original segmentation algorithm. Second, is to evaluate the computational cost, at least, in the segmenting process.

6.4.1 Evaluation framework specification

Objectives: The objective of this validation test is to measure the improvement of the proposal in the main objective, i.e. the segmentation results for the different object scales appearing

Fig. 6.4. Initial seeds per local feature scale value.

in the image. Additionally, depending on the characteristics of the image and/or the target application, it can also result in a reduction of the computational cost.

Dataset: We define an experiment-oriented database. The database requirements are the following:

1. We need a database including a high precision segmentation ground-truth.
2. We need a database containing target and background categories.
3. We need a database including targets with different sizes.

We use as baseline the Davis Video Segmentation database [Perazzi et al., 2016]. As this is only a validation test, we selected 10 images from the database which meet the database requirements. Figure 6.5 shows the selected images.

Methodology-metric: We evaluate both aspects separately.

1. Boundary adherence evaluation. The segmentation process is similar in both techniques. To obtain the most reliable evaluation, we propose the following methodology. We first run the original SLIC segmentation algorithm with 5 different spatial constraints -scales-. Then, we run the proposed LF-SLIC. For each image, for the target in the image, we obtain the contour accuracy F. This quality metric is defined in the database article [Perazzi et al., 2016]. It reflects, at pixel level, how the boundaries of the target object are segmented. Therefore, the smallest scales will usually obtain the best results. We will include the number of regions generated at each scale for a fair comparison. We would expect our algorithm to perform for each target almost as

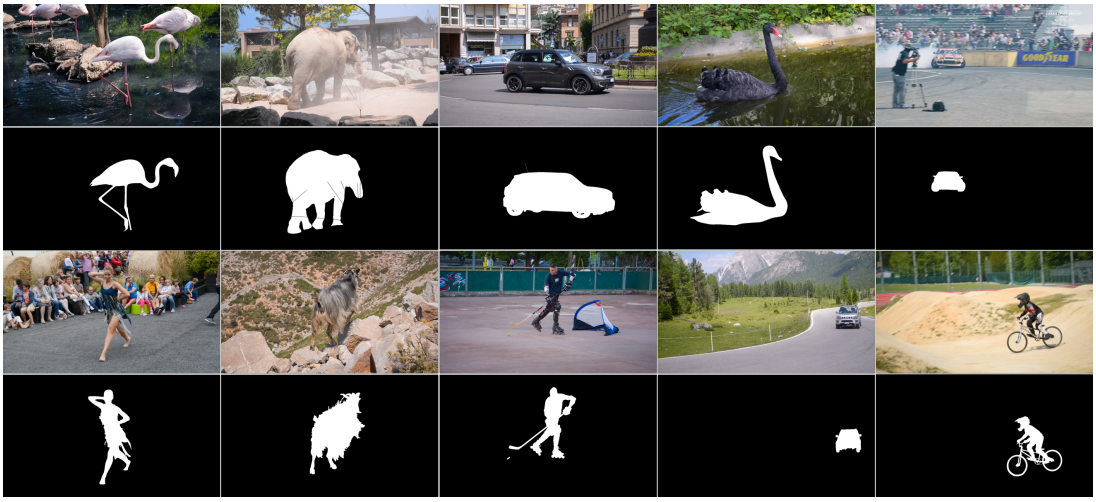


Fig. 6.5. LF-SLIC validation database. It contains ten images of the first ten categories of the Davis Video Segmentation database [Perazzi et al., 2016]. Top row, images 1 to 5. Third row, images 6 to 10. Below each image, we include the segmentation mask.

good as the best of the 5 different SLIC segmentations with a remarkable reduction in the number of regions generated (with further impacts on the computational cost).

2. Computational costs evaluation. We define the computing capabilities: *Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz, 4 Cores, 32 Gb RAM*. We run the superpixel segmentation original method SLIC and the proposed LF-SLIC. Both techniques are implemented in Matlab for the evaluation. We use the same spatial constraints for the SLIC method and for the original scale of the proposal. We compare two aspects: processing time and resulting segmentation size.

6.4.2 Results

Results in Table 6.2 shows a remarkable performance of the proposal. To properly evaluate the results, we must consider also the number of regions generated. Using that information, we can do two straight comparisons. First the LF-SLIC against the scales σ_1 and σ_2 –the ones with a similar number of generated regions–. The proposal obtains an average improvement of 91% against scale σ_1, and and average improvement of 11% against scale σ_2. Second, we compare the proposal against the scale with the best F accuracy –scale σ_4–. That scale obtains a 0.91 average F accuracy, and our proposal 0.89. However, if we compare the number of regions, the LF-SLIC segmentation algorithm obtains these results generating 10 times less regions.

Results in Table 6.2 shows an improvement in the computational cost. As expected, the improvement depends on the image. In those images where there are big homogeneous areas, e.g. images 08, 09 and 10, the results are much better for SLIC. In those images where clutter

F accuracy	SLIC					LF-SLIC
	σ_0	σ_1	σ_2	σ_3	σ_4	
image01	0.4197	0.4319	0.8803	0.9065	0.9947	0.9433
image02	0.2065	0.5298	0.7475	0.7859	0.8077	0.7825
image03	0.4760	0.4968	0.7282	0.8443	0.9238	0.8994
image04	0.6272	0.6751	0.8939	0.9154	0.9629	0.9269
image05	0.1643	0.2178	0.8519	0.9072	0.9139	0.9102
image06	0.3954	0.4067	0.7365	0.8185	0.9367	0.8864
image07	0.2636	0.2757	0.6757	0.7445	0.8026	0.7971
image08	0.3850	0.5627	0.8747	0.9499	0.9897	0.9825
image09	0.6113	0.6536	0.7849	0.7962	0.8152	0.8025
image10	0.6019	0.4031	0.8233	0.9839	0.9887	0.9782

# Regions	SLIC					LF-SLIC
	σ_0	σ_1	σ_2	σ_3	σ_4	
image01	45	105	512	992	3680	282
image02	43	103	511	999	3814	391
image03	44	103	492	980	3787	642
image04	45	104	505	999	3830	346
image05	43	101	493	996	3871	239
image06	44	104	487	997	3887	412
image07	45	105	502	1010	3919	483
image08	43	98	500	963	3780	270
image09	45	103	504	978	3801	244
image10	45	104	506	963	3817	253

Table 6.1: Results in terms of F accuracy and #regions results. The top Table presents the F accuracy results for each image. The bottom one presents the #regions results for each image. We evaluate 5 different runs of the SLIC method -different scales- and one of the LF-SLIC.

Criteria	Processing time (s)		Result size (#regions)	
	SLIC	LF-SLIC	SLIC	LF-SLIC
image01	0.973	**0.419**	512	**282**
image02	0.958	**0.708**	511	**391**
image03	**0.960**	1.258	**492**	642
image04	1.098	**0.625**	505	**346**
image05	1.063	**0.377**	493	**239**
image06	1.011	**0.832**	487	**412**
image07	1.091	**0.952**	502	**483**
image08	0.977	**0.404**	500	**270**
image09	0.991	**0.379**	504	**244**
image10	1.019	**0.382**	506	**253**

Table 6.2: Computational cost comparison. SLIC and LF-SLIC are compared for each image analyzing the processing time and the # of regions generated.

appears, results are not that better. The proposal's results for image 03 for example are worse than the original. The reason is the cluttered background with a lot of corners generating LF detections. The real computational cost improvement will be noticeable in those applications that process the resulting regions.

6.5 Conclusions

The proposed combination region-segmentation vs LF, specified in the LF-SLIC method, overcomes SLIC original limitations in terms of multiscale support. The proposal presents remarkable segmentation results with a reduction of the oversegmentation. This is achieved using LF detections to define those areas where the oversegmentation is required, and those where not. The benefits of LF-SLIC have been shown in this chapter but will be extended in the following chapter in a real scenario applications. Essentially, this chapter proposes a method to partially validate the thesis main hypothesis.

.

Chapter 7

LF-SLIC validation via skin lesion segmentation application

In this chapter we present a successful application of the LF-SLIC method. We motivate the application selection. We present the application where the method is the core of the solution. Finally, we present a further application that may be built combining the two proposed schemas in this thesis. As a result, from the application evaluations we validate the proposed technique and thus the second part of the thesis main hypothesis. This chapter is built on the work developed for the published international journal article: "Accurate segmentation and registration of skin lesion images to evaluate lesion change", [Navarro et al., 2018a].

7.1 Application selection

The proposed solution, LF-SLIC, has been validated from a functional point of view in Chapter 6. However, it is in real conditions testing where the solutions, and thus the hypothesis, can be considered proven. The motivations of LF-SLIC from the capabilities point of view are clear. However, in the application level the objective is the following: provide the superpixels new capabilities so we can take advantage of their properties in new applications.

We have used the same criteria to select an LF-SLIC application than that used for the SP-SIFT validation in Chapter 5. These criteria are the following: 1) An application where superpixels are rarely used due to their drawbacks; 2) An application subject to extensive research, so we can compare our proposal with top quality algorithms; and 3) An impactful application, where the improvement can be seized and future researchers can take advantage of our work.

After considering a number of applications, we selected **skin lesion segmentation** which fulfils the aforementioned criteria:

- Challenge: skin lesions presents a great variability in scale. The boundaries of the lesions are sometimes highly textured. Additionally, capturing this slight variation in the boundaries is critical for further diagnosis. This problem needs a solution with the capability of precisely segmenting regions in an image at different scales, which is exactly the challenge that our proposal was intending to solve.

- Scientifically active application: skin lesion segmentation has been recently pushed forward with the apparition of the ISIC challenge [1]. The first challenge was ISIC 2017. They provided around 2.000 images belonging to three categories: melanoma, seborrheic keratosis and benign nevi. ISIC 2019 has more than 25.000 images belonging to 9 categories of skin lesion.

- Impact: health applications have a major impact when compared with other tasks. Skin cancer is the most prevalent form of cancer in the United States, with 5 million cases occurring annually. Lesion segmentation is required to develop automatic cancer detection tools. Early diagnosis is directly related to survival ratios.

Once the task is defined, we proposed the following application.

1) LF-SLIC as the core of an unsupervised skin lesion segmentation application. We propose a lesion segmentation solution based on the region segmentation capabilities of the LF-SLIC segmentation algorithm. Despite the challenging process of creating an algorithm to compete with the state of the art, this strategy allows us to validate the real potential of the proposed method.

2) SP-SIFT on top of LF-SLIC for later image registration. We propose a combination of our proposals. The former needs support regions to perform, and the latter provides region support for further applications. Despite the challenging process of combining the two proposed techniques, this strategy allows to validate how far we can push these techniques' capabilities.

7.2 LF-SLIC application: accurate segmentation and registration of skin lesion images to evaluate lesion change

7.2.1 Background on skin lesion segmentation

Due to the widespread unavailability of equipment and qualified human resources required to screen every patient, there is a need for an automated system to assess skin lesions and classify them into melanoma, non-melanoma and benign.

This application, apart from intending to validate the thesis hypothesis, presents contributions to the state-of-the-art in this direction.

[1] https://challenge2019.isic-archive.com/

Dermoscopy or Epiluminescence Microscopy (ELM) is a noninvasive imaging technique that helps diagnose skin lesions. ELM allows visualization of the subsurface structures of the skin revealing lesion details in colors and textures.

ELM improves the detection rate of skin lesions with respect to naked eye inspection, in which the highest accuracy is around 60% [Ma and Tavares, 2015]. Nevertheless, diagnostic accuracy using ELM largely depends on the dermatologist's experience. Well-trained generalist computer-aided diagnosis (CAD) systems are designed to reduce this dependency. CAD systems may also be used to monitor benign skin lesions in order to prevent their evolution to malignant lessons. Generally, a CAD system is composed of three major stages: *image segmentation, feature extraction, and classification.*

Image segmentation is used to locate the boundary between the lesion area and the surrounding skin. Obtaining an accurate segmentation of the lesion is important, especially to provide low error rates prior to later quantification of the shape, border and size cues of the skin lesion [Celebi et al., 2009]. In general, the segmentation process aims at the spatial discrimination of sets of inter-related pixels in a region of interest (ROI) to facilitate the detection of spatial transitions between these sets. Reported skin lesion segmentation methods are based on: edge extraction, image thresholding, region segmentation, artificial intelligence or active contours.

Edge based techniques [Barcelos and Pires, 2009; Chung and Sapiro, 2000] are based on information about the image edges; more specifically, they search for abrupt changes in the intensity of neighboring image pixels.

The segmentation process may also depend on similarity criteria, such as similar grey levels, colors or textures.

Thresholding[Garnavi et al., 2011] and region-based [Silveira et al., 2009] segmentation are examples of methods that use similarity criteria to identify skin lesions in images.

Techniques based on artificial intelligence (AI) [Yu et al., 2016] classify pixels as belonging to the lesion or to the background of the images. Neural networks, evolutionary computation and fuzzy logic are some examples of these techniques.

Algorithms based on active contours are also used for segmenting skin lesion images [Zhou et al., 2013]. In these algorithms, the initial curves evolve towards the boundaries of the lesion through appropriate automatic deformation.

Feature extraction plays a major role in automatic skin lesion diagnosis. In human-driven analysis of skin lesions, there are widely accepted templates to evaluate the evidence of particular lesions. For instance, dermatologists have created the ABCDE rule for Melanoma lesions. Melanomas tend to be Asymmetric, have an irregular Border, present uneven Color distributions, their Diameter is greater than 6 mm and they Evolve in size, shape and color.

The (E)volve feature is a key element in the diagnosis of pigmented lesions. Its extraction is based on a prior image registration stage. Therefore, image registration is a critical task and an

area that has been widely studied. Image registration can be done at full-body level, i.e. full-body images are registered to detect the apparition of new moles or the growth of pre-existing ones [Mcgregor, 1998]. Image registration can also be done at the level of individual single skin lesions [Anagnostopoulos et al., 2013]. Skin lesions are registered with millimetric precision so even the smallest changes in the lesion can be observed. The main techniques tend to rely on points matching [Perednia and White, 1992] or regions [Huang and Bergstresser, 2007]. Some solutions include a prior skin lesion segmentation process [Maglogiannis, 2003].

Advances in the feature extraction stage in CAD systems have been focused on the automatic extraction of these cues. The spatial pixel area extracted from the segmentation process has been analyzed to derive asymmetry, shape, border and diameter cues (ABCD rule) [Tsao et al., 2015].

Nevertheless, the automatic extraction of these cues is problematic mainly due to inaccuracies at the segmentation stage and to the complexity in registering images of a skin lesion taken at different times.

Classification consists in recognizing and interpreting the information on the pigmented skin lesions based on the cues extracted.

7.2.2 Skin lesions segmentation

We present here an application of the proposed LF-SLIC segmentation algorithm to ELM images of skin lesions. For the success of the application, we define two premises:

- Images capture a single lesion.
- The lesion is fully contained in the image.

These requirements could, however, be eliminated in future. The segmentation process is divided in two sequential stages: LF-SLIC region labeling, and Artifact removal.

7.2.2.1 LF-SLIC region labeling via spatial continuity classification

The image ψ is segmented into a set of LF-SLIC superpixels $\Omega_j, j = 1 \dots J$, where J is the number of superpixels.

Two superpixels, Ω_j and $\Omega_{j'}$, are neighbors if at least one of the pixels in Ω_j is 8-connected with a pixel in $\Omega_{j'}$. Let $\overline{bw}$ be the bandwidth of this partition, defined as the largest 5-dimensional distance vector –evaluating position and RGB color– between the centers of any two neighboring superpixels.

The final objective is to obtain two disjoint sets of superpixels: a subset of superpixels classified as non-lesion $N = \{\Omega_p, p\epsilon[1..J]\}, |N| = P$; and a subset of lesion superpixels $L =$

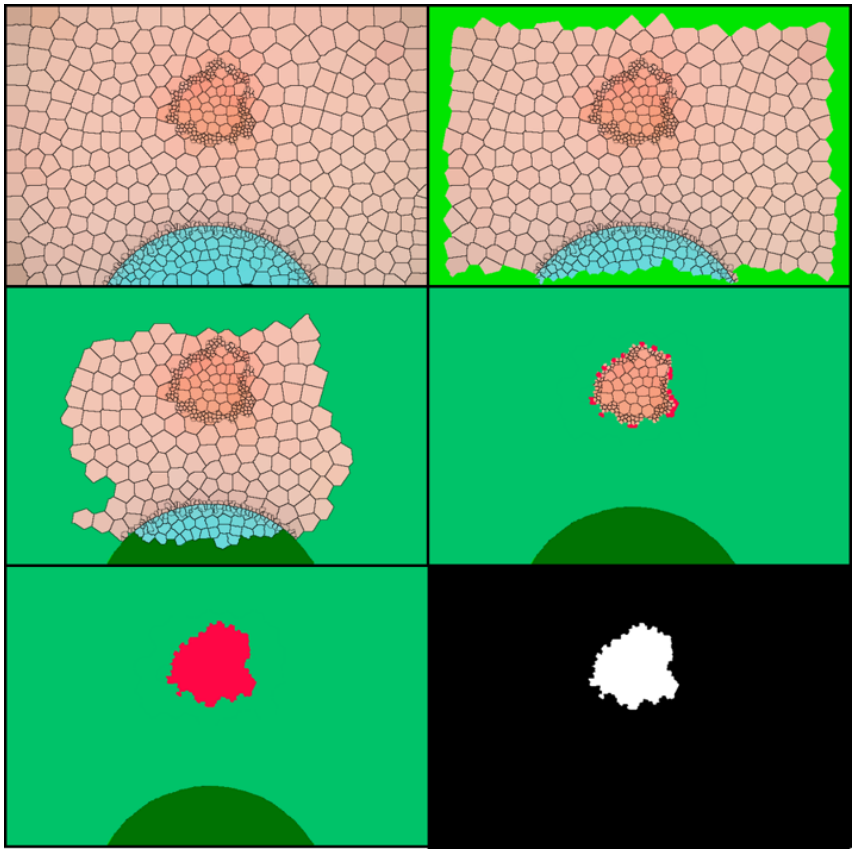


Fig. 7.1. LF-SLIC labeling process. The top left image shows the LF-SLIC superpixels segmentation. The top right image shows the N^0 set of superpixels in light green. The mid left image shows an iteration t, where different green areas indicate different clusters formed in theN^t set. The mid-right image shows in red the superpixels classified into the L^t set for a later iteration. The bottom row shows the final classification: the left image describes the final clusters (green for the N set and red for the L one) while the right one depicts the final segmentation mask.

$\{\Omega_q, q\epsilon[1..J]\}, |L| = Q$, where $P + Q = J$. For this purpose, a greedy labelling scheme with connectivity restrictions is proposed.

First, under the assumption that the lesion is fully contained in the image, all superpixels that are 8-connected to the image boundary are assigned to the N set –see Figure 7.1 top image–, creating an initial estimation of non-lesion superpixels, N^0, and the complementary initial set of lesion superpixels,L^0.

Then, superpixels in the N^0 set are grouped into regions using a conservative mean-shift approach [Comaniciu and Meer, 2002] with a bandwidth $\overline{bw_{MS}} = \overline{bw}$. This process merges superpixels in N^0 into regions $\{R_1 \dots R_m \dots R_M\}$, each containing a subset $\{\Omega_{p,m}\}$ of the N^0 superpixels. Due to this conservative grouping, the set of colors $\{\overline{c}(\Omega_{p,m})\}$ of the superpixels in every region can be assumed to define a close-to-Gaussian-distribution.

Under this assumption, superpixels in L^0 are reclassified by evaluating their likelihood to be part of any of their 8-connected regions in the set $\{R_1 \dots R_m \dots R_M\}$. For this purpose, for a superpixel in L^0 with color $\overline{c}(\Omega_q)$ that is connected to region R_m, a Grubbs' test is used to determine whether the superpixel is an outlier of the color distribution inside R_m:

$$G = \frac{|E\left[\overline{c}(\Omega_{p,m})\right] - \overline{c}(\Omega_q)|_2}{\sigma\left[\overline{c}(\Omega_{p,m})\right]} \tag{7.1}$$

with $E\left[\overline{c}(\Omega_{p,m})\right]$ as the mean vector of the colors of the superpixels in R_m and $\sigma\left[\overline{c}(\Omega_{p,m})\right]$, its standard deviation. The hypothesis of Ω_q being part of R_m is accepted at significant level α, fixed in the experimental parameter setup. If:

$$G \leq \frac{M-1}{\sqrt{M}} \sqrt{\frac{M-1}{M-2-ts^2_{\alpha/2M,M-2}}} \tag{7.2}$$

where M is the number of superpixels in the set, and ts the Student's t-distribution.

Reclassified superpixels are removed from L^0 and assigned to N^0. This process is repeated for any superpixel in L^0 which is a neighbor of at least one region $\{R_1 \dots R_m \dots R_M\}$, creating two new sets L^1 and N^1.

The whole process is repeated until at a given iteration, say t, no further reassignments are performed. The sets at this iteration $N = N^t$ and $L = L^t$ define a tight-to-boundaries segmentation of the skin lesion –see Figure 7.1 for iteration examples–. However, artifacts in the image can also be segmented, so we propose an artifact removal method.

7.2.2.2 Artifact removal

The first step for the successful removal of artifacts is to define an artifact precisely. According to state-of-the-art reports, artifacts in ELM images mainly consist of hair and air bubbles.

These artifacts clearly differ from the skin lesions. Skin lesion boundaries show irregular

shapes, and present smooth transitions with the surrounding skin; on the contrary, the artifacts identified show contours that contrast greatly with the surrounding skin and also very regular shapes: straight lines for the hairs and circles for the bubbles.

There are many established approaches to detect pre-defined and highly contrasted shapes. For this purpose, we propose to use the well-known Hough Transform (HT) [Pao et al., 1992], a voting scheme that obtains highly robust detection results in these situations. We apply the HT to detect pixels belonging to lines, circles or ellipses in the segmented image. Superpixels containing pixels voted as lines, circles or ellipses are re classified into the N set.

7.2.3 Skin lesions registration to evaluate change

The second contribution of this work is to measure the evolution of a skin lesion, given two images capturing different stages of the lesion, a crucial criterion for diagnosis.

The result of the proposed segmentation is a precise image of the isolated skin lesion, which allows for the extraction, for instance, of the ABCD cues to further classify the lesion. If we have two images of the same lesion captured on different days, we could measure the change or evolution (E) in the ABCD cues. However, for this process to be reliable and effective, both images should show the lesion with a comparable scale, orientation and point of view; that is, an image registration process should first be performed.

7.2.3.1 Image registration with the SP-SIFT feature

Registration requires identifying the same feature points in the two images to perform proper image alignment.

State-of-the-art algorithms for skin lesion registration face the problem of aligning reference cues that may have suffered remarkable changes (evolution).

We propose to use the SP-SIFT technique to detect and describe feature points in both images first, so that the evolution of the skin does not corrupt the characterization of the feature points. Detected LF are used to establish matching points between these two images. These matches define a geometric transform (in this case, an homography) between the pair of images. We use the transformation to align both images. An example of the image alignment process is depicted in Figure 7.2.

7.2.3.2 Evaluation of the lesion change

In this paper we do not explore the extraction of cues to characterize skin lesions. Instead, we focus on obtaining a precise segmentation in order to extract the desired cues more accurately, and on registering skin lesion images to allow comparison of cues extracted at different times and then evaluate lesion change.

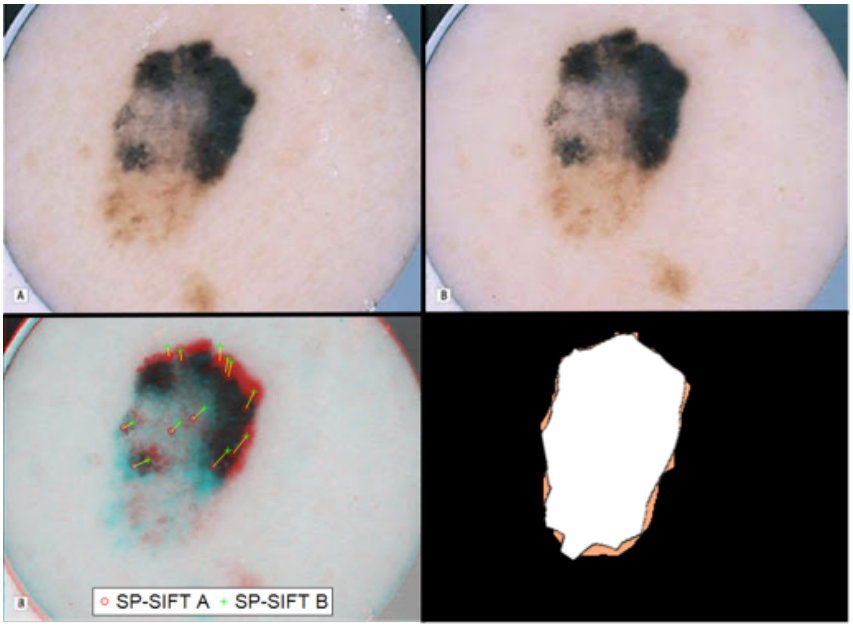


Fig. 7.2. Skin lesion registration and size evolution. The top row shows the first (A) and second (B) skin lesion images. The bottom left image shows the matched SP-SIFT feature points between both input images. The bottom right image shows the segmentation masks aligned or registered for easy use in size comparison.

In order to illustrate and demonstrate the potential of our proposal, we present in the next section results on the evolution of the size of the lesion, one of the main characterization cues. Variations could also be obtained for color, boundaries or asymmetry; however, this falls outside the scope of this work.

We use the image registration technique to align both skin lesions and their segmentations. We compare the segmented areas and calculate a pixel level difference. The scale of the images is known, so we can map pixels to millimeters and provide the size-feature evolution in a comprehensive metric.

7.2.4 Experimental results

We present here the results of a comparative analysis of the proposed segmentation method, using the benchmark proposed in the scope of the ISIC 2017 challenge. Additionally, we evaluate the proposed image registration method against a modified version of the ISIC 2017 test set. Finally, we show how these methods allow a precise evaluation of the variation in the diameter of a skin lesion.

7.2.4.1 Evaluation of the proposed segmentation method

Data analyzed: We have arranged the data according to the ISIC 2017 evaluation framework:

- Training data: 2000 dermoscopic images and their respective 2000 binary ground-truth masks.
- Validation data: 150 dermoscopic images and their respective 150 binary ground-truth masks.
- Test data: 600 dermoscopic images and their respective 600 binary ground-truth masks.

The training data are used to set the algorithm parameters; the validation data are used to assess the setup; the test data are used to evaluate the proposed algorithm and to perform the comparison with alternative state-of-the-art algorithms.

Evaluation measures: We first select the Jaccard Index (J), which is one of the most widely used metrics to evaluate segmentation methods, and the one used in the ISIC 2017 challenge. J is also known as intersection over union. It is defined as the ratio$J(A,B) = |A \bigcap B|/|A \bigcup B|$, where A and B are two binary masks; and it provides a normalized measure, the higher the better, of the overall performance of a segmentation method. We complement this indicator with the Dice coefficient (S), also widely used to evaluate the similarity between two binary masks. S is usually considered to be a semi metric version of $J : S(A,B) = |A \bigcap B|/(|A| + |B|)$. Additionally, as segmentation can be viewed as a pixel classification task, performance can also be measured by a classification quality indicator. We used Accuracy: $ACC = TP + TN/(TP + TN + FP + FN)$.

System setup: Default parameters are used for the LF-SLIC, SIFT and SP-SIFT methods. The Mean-Shift bandwidth is set according to the LF-SLIC result. Hence, the training and validation data are just used to set the value of the significant level α associated to the Grubbs' test. For this purpose, we have obtained ACC values for the range $\alpha\epsilon[0.7, 0.99]$. We selected as a trade-off value the one that returns the highest value in both sets. In the experiments, this value was $\alpha = 0.91$ achieving $ACC = 0.998$ in both validation and training sets.

Quantitative results: The proposed method is compared –see Table 7.1 – Proposed-1– to the Top 5 algorithms in the ISIC 2017 Challenge; the Dice Coefficient and the Accuracy are also included. To assess our method's performance better, we also include our results –see Table 7.1 – Proposed-2– previously removing from the dataset those images that do not fulfill our assumptions (i.e. images where the skin lesion is not fully contained in the image). To evaluate further the operational range of the methods compared, Figure 7.5

Reference	Jaccard Index	Dice Coefficient	Accuracy
Top 1	0.765	0.849	0.934
Top 2	0.762	0.847	0.932
Top 3	0.76	0.844	0.934
Top 4	0.758	0.842	0.934
Top 5	0.754	0.839	0.934
Proposed-1	**0.769**	**0.854**	**0.955**
Proposed-2	**0.846**	**0.938**	**0.960**

Table 7.1: Segmentation results ISIC 2017 Challenge [Codella et al., 2018].

depicts box-plot diagrams of the Jaccard Index distribution: the vertical size of the box indicates results dispersion (standard deviation) and the horizontal lines represent average values; points outside the boxes are outliers.

7.2.4.2 Evaluation of the proposed lesion registration method

The aim of this experiment is to assess the effectiveness of the proposed registration method in the task of aligning skin lesion images. We compare the performance of the SP-SIFT technique in this task respect to two well-known feature detection-description algorithms: SIFT [Lowe, 2004] and SURF [Bay et al., 2006].

Data analyzed: In order to carry out a systematic evaluation, we use the ISIC 2017 test dataset as the set of initial skin lesion images (i.e., those corresponding to the initial lesion capture), and we then generate for each image in this test set, a new image simulating a capture in a different instant/conditions: we randomly generate one of the following modified images: a illumination change, a rotation or orientation change, a scale change, or a change in the point of view –see Figure 7.3–.

Evaluation measures: Each technique compared extracts LF from both the original image and each of the modified images and matches them to establish correspondences between the initial and the modified image. The quality of the correspondence is then evaluated in terms of average precision and recall: if the correspondence is correct, a true positive is declared (TP); if it is incorrect a false positive is declared (FP); if no correspondence is established, a false negative is declared (FN). Precision (P) and recall (R) of the matching process are then defined as $P = TP/(TP + FP)$ and $R = TP/(tp + fn)$.

Quantitative results: Figure 7.4 includes the results obtained for the three techniques on the modified version of ISIC 2017 test dataset in terms of average precision and recall. Results are given for each image modification.

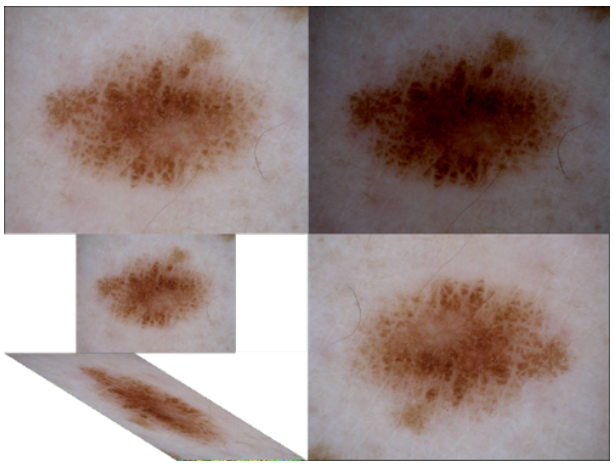


Fig. 7.3. Example of image distortion applied to the ISIC 2017 segmentation test set. Firs row original image (left) and light change (right). Second row, scale change (top left), viewpoint change (bottom right) and orientation change (right).

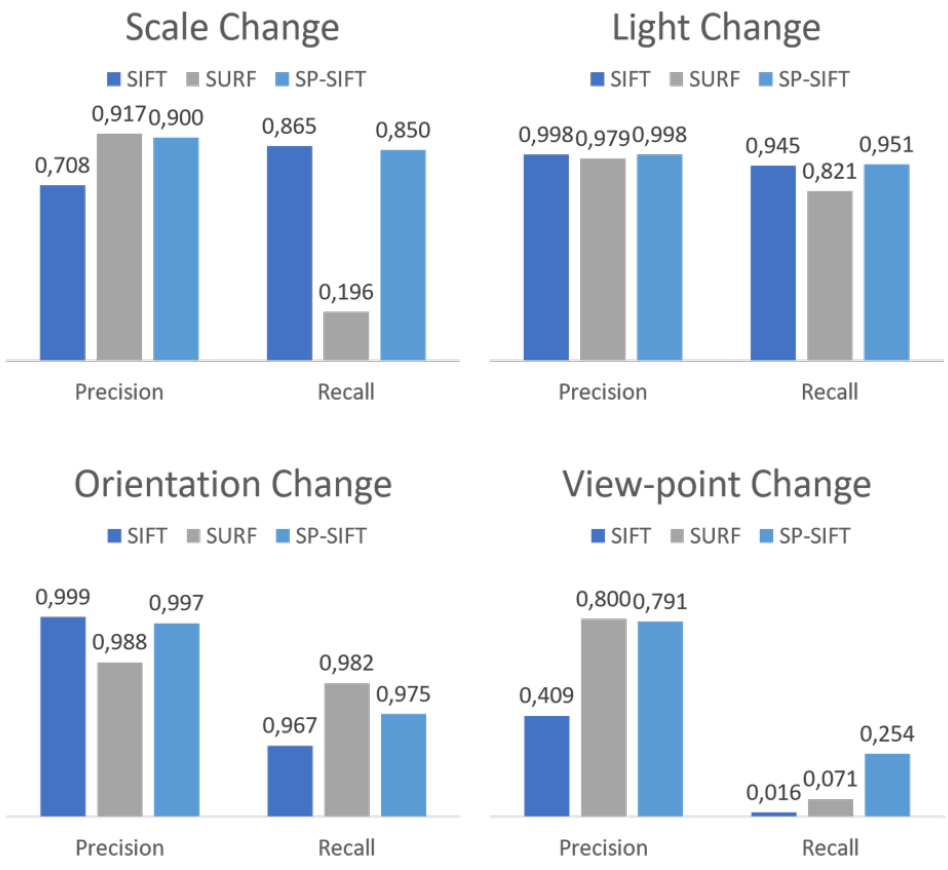


Fig. 7.4. Precision and Recall matching results for modifications of the images in the ISIC 2017 test set.

Category	Ground truth diameter evolution (mm)	Estimated diameter evolution (mm)	ε (mm)	Average matched LF
No Change	0	0.01	0.01	64.32
Short time	1.28	1.46	0.18	12.93
Mid/large time	4.51	5.15	0.60	3.23
Overall	1.63	1.86	0.23	26.56

Table 7.2: SP-SIFT image registration and diameter evolution results.

7.2.4.3 Case study: Assessing the evolution of the lesion diameter

In this experiment the objective is to present a potential application of the image registration process: measuring the evolution of the lesion's diameter.

Data analyzed: For this experiment, we use a subset of the [Menzies et al., 2001] dataset. This subset contains 10 pairs of images from 10 different patients. Temporal distance between images of the same patient ranges between a few days (6) and a few months (4.5). Each pair of images has associated ground-truth information indicating the diameter variation between the first and the second image.

Evaluation measures: We perform the evaluation based on two criteria. The average number of correctly matched points between the two temporally spaced samples and the error in mm (ε) between the predicted and the annotated diameter change. Note that the image registration process, i.e. the homography estimation, requires at least three matched points.

Quantitative results: Table 7.2 shows the average results of the evaluation. To evaluate the capabilities of the method better with respect to time variation, images are grouped into three categories:

- No change: the skin lesion analyzed suffered no change between the first and the second picture.
- Short time: the time elapsed between the first and second images is less than 2 weeks. Changes are expected to be small.
- Medium to long time; the time elapsed between the first and the second image is more than 2 weeks. Changes are expected to be bigger than in the short time category.

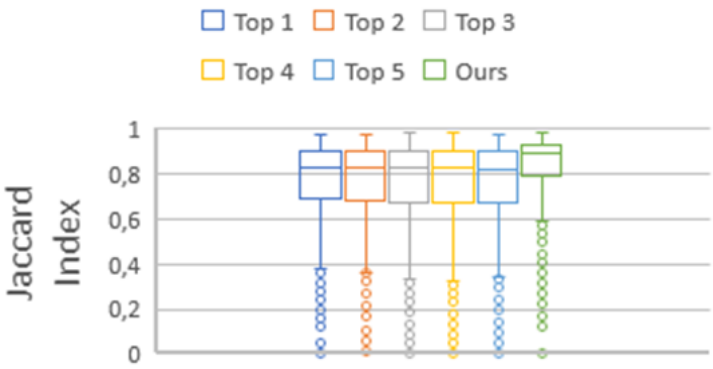


Fig. 7.5. Distribution of the Jaccard Index for all the images in the test set of the ISIC 2017 segmentation challenge. See text for discussion.

7.2.5 Discussion of the proposed application experiment

7.2.5.1 Proposal for the lesion segmentation

In the segmentation stage, we extracted referenced results of state-of-the-art methods from the ISIC 2017 skin lesion segmentation challenge. Top-ranked algorithms present Jaccard Indexes ranging from 0.765 to 0.754, all very close –see Table 7.1–.

The proposed segmentation method yields a Jaccard Index of 0.769, outperforming the other approaches. Besides, the proposed method also performs better in terms of the Dice Coefficient and Classification Accuracy. Results are obtained using the whole test set, including images that do not meet the method's prerequisite of having the skin lesion fully contained in the image. For a deeper understanding of the segmentation results, we include a box plot graphic in Figure 7.5.

The proposed method also outperforms the other methods by yielding a lower deviation, i.e. its operation is more stable for more images in the set. However, the distribution of the Jaccard Index achieved by the proposed method presents a higher number of outliers than the other methods. These outliers are basically the images which do not meet the prerequisite. If these images are removed, results improve up to 0.846 in Jaccard Index terms, 10.56% better than the top approach in the challenge –see Table 7.1–.

Results of the proposed approach –and of all the other approaches evaluated– are biased by the annotated ground-truth. Despite the high quality of the dataset, and the amount of data provided, the annotation of skin lesions is a subjective task. This can be observed in the failure cases presented in Figure 7.6. The ground-truth annotations of the images in the two first columns are not tight to the lesion itself, but rather include a roughly affected spatial area around it which substantially differs from the proposed segmentation, which is tighter to the lesion. Differently, the third column depicts an example of an annotation mistake, in which the

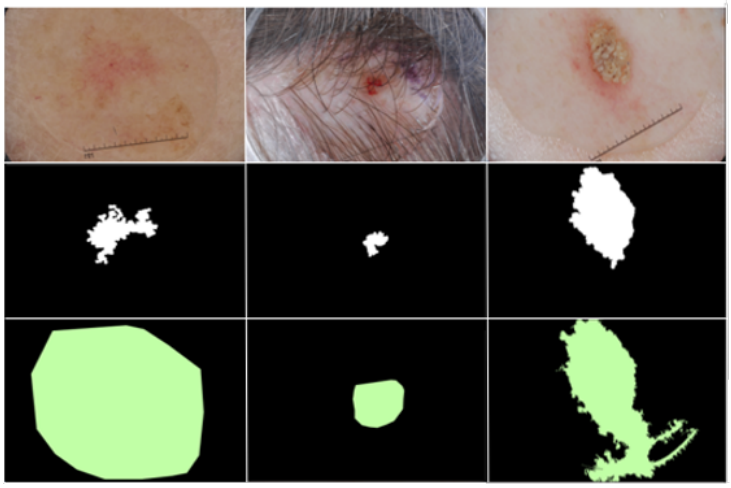


Fig. 7.6. Failure cases (three of the outliers in the Jaccard Index distribution presented in Figure 7.5). First row, dermoscopic images. Second row, segmentation results obtained with the proposed method. Third row, ground truth segmentation.

ruler is included in the ground-truth mask.

Despite the good results obtained, there is room for improvement. Superpixel segmentation provides a robust tool for skin lesion segmentation. However, the accuracy of the segmentation on the lesion boundaries is biased by the superpixels' sizes and shapes. Despite the high accuracy achieved by the LF SLIC, it can be improved by operating at pixel level.

7.2.5.2 Proposal for the lesion registration and evolution assessment

Although it is a key stage for the extraction of feature evolution, to the best of our knowledge, there is no prior study dealing with skin lesion registration. We present a comparison between state-of-the-art LF, as they have been shown to be successful tools for image registration in other fields.

According to Figure 7.4, the SP-SIFT descriptor used for describing the superpixels obtained by the LF SLIC segmentation, yields better results than the SIFT and SURF techniques. Light changes are well handled by both the proposed SP-SIFT scheme and the SIFT LF. The scale changes affect the SIFT LF slightly, but the proposed version of SP-SIFT is robust to these changes due to the tightness of the description supports. Finally, whereas geometric changes in terms of image orientation are handled well by all three methods, affine transformations or point of-view changes are still challenging. Despite the proposed version of SP-SIFT yielding a recall 8.48% and 46.37% better than SIFT and SURF, its results can be still substantially improved.

The image registration is presented here as a tool to facilitate the extraction of feature evolution (E). Dermatologists agree on the relevance of the cues evolution over time in order to

detect potentially malignant lesions. Whereas there are some studies that describe strategies to extract this feature, the complexity of the process hinders the existence of robust automatic approaches and of state-of-the-art evaluations.

In this proposed application, the potential of image registration is exemplified by evaluating the variation in the diameter of 10 different skin lesions. The results obtained –see Table 7.2– indicate that there is an average error of 0.23 mm between the estimated and the real evolution of lesion diameters. Considering that the critical diameter of a skin lesion is 6 mm, the error represents a deviation of 0.04% of this magnitude. However, results also suggest that accuracy degrades with the time elapsing between lesion samples, suggesting that a continuous observation of the lesion will be required for effective assessment of the lesion's evolution. The downgrading can be explained by the decrease in the average number of matched LF. For large time lapses, the average number of correctly matched LF is close to three, the minimum number required for image registration. In these situations, image registration may be driven by incorrectly matched LF.

7.3 Discussion on the application of the LF-SLIC region segmentation algorithm

The two main objectives of this application were to evaluate the success of the LF-SLIC algorithm for the accurate segmentation of skin lesions, and an application of the SP-SIFT feature for the accurate registration of two images of the same skin lesion.

Moreover, these algorithms operate together to achieve a more challenging objective: a precise segmentation mask enables the extraction of precise cues characterizing the skin lesion; precise registration further allows reliable measurement of the evolution of such cues, which is also in a major contribution of this application.

We consider successful the proposed segmentation algorithm, LF-SLIC, and its combination with the robust artifact removal technique. Results demonstrate that they achieve top state-of-the-art results with the dataset provided by the ISIC 2017 skin lesion segmentation challenge.

We also consider a success the proposed technique for the registration of skin lesion images. The proposal uses a feature point detection and description technique, the SP-SIFT. The experimental results show that the proposal can perform the skin lesion registration under different capture conditions and lesion stages.

Finally, the combination of these techniques, an accurate segmentation and a reliable image registration, paves the road for the precise computation of features' evolution and automatic skin lesion classification, and further applications of those techniques in other pathologies.

Part V

Conclusions

Chapter 8

Achievements, conclusions and future work

8.1 Summary of achievements and main conclusions

In this thesis, we considered the problem of local features detectors/descriptors and region segmentation algorithms for image characterization, aiming to find an effective way to combine them for improving their individual capabilities.

To start, we reviewed both techniques and outlined the main properties to be considered when developing new methods and analyzing existing ones (Chapter 2). We then proposed (Chapter 3) a hypothesis for combining their capabilities yielding a pair of combination strategies, SP-SIFT and LF-SLIC, that lead to effective generalization.

The first strategy, SP-SIFT, is focused on the enhancement of local features, specifically of the SIFT local feature [Lowe, 2004], using region segmentation algorithms—SLIC superpixels [Achanta et al., 2012]—to guide the image information description process (Chapter 4).

The second, LF-SLIC, is focused on the enhancement of region segmentation techniques, specifically SLIC superpixels, using local feature detectors—SIFT local feature detector—to guide the scale in the segmentation process (Chapter 6).

To evaluate the hypothetical advantages of these strategies, we considered two of different applications. First, we showed how to use the SP-SIFT local feature as a cue for improving state-of-the-art tracking solutions by the background-inhibition nature of the SP-SIFT strategy in the description (Chapter 5). Then, in the context of region segmentation, we presented an approach for using the LF-SLIC region segmentation algorithm in medical imaging, specifically for pigmented skin lesions segmentation. Obtained results suggest that the LF-SLIC schema benefits for the SIFT-driven superpixel initialization (Chapter 7).

* * *

The combination schemas proposed are examples of powerful techniques resulting from the combination of local features and region segmentation algorithms, enhancing baseline methods and generally improving the state-of-the-art for the studied computer vision applications. The potential application fields that may benefit from the achieved capabilities are worth studied. Remarkably, the combination schema not only allows the methods to work together, but also aims to enhance their strengths and minimize their weaknesses. In our opinion, future researchers will keep on combining local features and region segmentation algorithms with different properties, hopefully yielding more powerful and versatile solutions adapted to their target applications.

8.2 Discussion of open questions and future work

As is the case with many academic endeavors, the research presented here opens many more questions than answers. Below, we discuss some of the opened questions and future directions that we believe may be followed to answer them.

Part II The objective of this part is to present the hypothesis, and the two statements that will be evaluated to prove it truth. To do so, in Chapter 2 we discuss the state-of-the-art of both local features and region segmentation algorithms. We analyze their properties and select the two techniques that we consider to be the best suited for the thesis development: SIFT local feature [Lowe, 2004] and SLIC superpixels [Achanta et al., 2012].

In Chapter 3, based on the analyzed properties, we present the thesis hypothesis: **"The combination of LF —based on local descriptions— and region segmentation algorithms —based on spatial and color relationships— results in a new family of image features with a wider application field thanks to their complementary strengths: discriminative capacity and robustness to variations in the local information"**. We divide the hypothesis in two statements to prove it truth. First, the use of a region segmentation algorithms is beneficial for improving the description process of the local features. Second, the use of local features detectors is advantageous for improving the region segmentation results.

For the future work, it would be interesting to consider the possibility of selecting different methods for their combination. The idea is to extend the proposed strategy to different local features and regions segmentation algorithms. In fact, SIFT and SLIC were chosen, not only for being top in performance in their respective categories, but for being also considered standards. Similar techniques should be easily integrated in the same combination schema. Specifically, we propose two initial combinations. First, it can be interesting to see how the combination schema affects to local features with circular description patterns. We suggest to create SP-DAISY or SP-GLOH. Second, it can be useful to evaluate learned local features detectors, as TILDE or LIFT, to initialize the region segmentation techniques.

Part III The objective of this part is to define a strategy that successfully combines local features and region segmentation algorithms in the way defined in the first statement. It is also an objective of this part to validate the proposal in real applications.

To this aim, in Chapter 4 we present the SP-SIFT local feature. A method that uses the superpixels segmentation of SLIC to isolate the local information that the SIFT feature will include in its description process. We also validate the characteristics of this new method. We first prove that it conserves the original capabilities of SIFT capabilities. Then, we prove that it provides additional capabilities as being more robust and discriminative in crowded scenarios.

In Chapter 5 we use the SP-SIFT in a computer vision application to prove its theoretical capabilities. We select tracking as the target application. We propose a two-step integration process of the local feature. The objective is to sequentially prove its performance, first without external factors and then as the core cue that drives the tracking. Therefore, we initially integrate the SP-SIFT local feature in a state-of-the-art tracker measuring the potential of the method without additional issues deriving from the tracking method. Then, we go a step further and propose a tracking algorithm with the SP-SIFT feature in the core, defining novel tracking strategies that are only possible thanks to the capabilities provided by SP-SIFT.

The results of both applications confirm the capabilities of the proposed technique as well as prove truth the first statement of the thesis for tracking applications.

For the future work, it will be interesting to evaluate the proposed method in alternative applications. We suggest to evaluate the impact of the feature in medical imaging applications, a hot and potentially impactful field in the last years.

Part IV The objective of this part is to define a strategy that successfully combines local features and region segmentation algorithms in the way defined in the second statement. As in Part III, it is also an objective of this part to validate this proposal in real applications.

In Chapter 6 we present the LF-SLIC region segmentation algorithm. A method that uses the SIFT's local features detector to improve the segmentation results of the SLIC superpixels, by obtaining a multi resolution superpixels technique without running segmentations at multiple scales. In the same Chapter we validate the characteristics of this new method. We first check if the segmentation performance is still comparable to SLIC's or better. Then, we compare also the benefits in efficiency, i.e. the saving in computational cost. Both evaluations were positive, i.e. support the statement, even if the major impact *was* expected to be noticed in the target applications.

In Chapter 7 we applied the LF-SLIC method in a computer vision task to prove its capabilities. We select skin lesion segmentation as the target application. We propose a solution to isolate the skin lessons in an automated fashion. The proposal includes the LF-SLIC in the core. The results obtained proved both the capabilities of the new proposal and the veracity of

the second statement for the target task.

For the future work, we consider interesting to evaluate the proposed method in different applications. We suggest to evaluate the method in automotive applications, e.g. line and vehicles detection, a recurrent research area which advance, albeit constant, is slower than expected.

Finally, we consider that both proposals, their validation and the results of the applications, proves truth the hypothesis of the thesis in the evaluated schemas.

Alternative combinations From this point, we think that it can also be interesting to evaluate the combination of techniques different from local features and region segmentation algorithms under the same schema. We consider that there will be only two restrictions in terms of the techniques that can be combined. First, related to the SP-SIFT-like combinations, the only restriction is to combine methods including information of a local environment to describe the image and methods with the capability of grouping image information with a certain level of semantic meaning. Second, related to the LF-SLIC-like combinations, the restriction is to use techniques identifying key points or areas in the image and, again, techniques grouping image information with certain a level of semantic meaning.

Specifically, we envision two initial approaches. First, semantic segmentation techniques [Kreso et al., 2017; Rota Bulò et al., 2018] can be tested to guide the local features description process. Semantic segmentation is defined as the task of clustering parts of images together which belong to the same object class. Second, it can be useful to evaluate the use of visual attention models [Zhao et al., 2015; Wang and Shen, 2017] to initialize region segmentation techniques. Visual attention methods attempt to determine the amount of attention steered towards various regions in an image by the human visual and cognitive systems.

Part VI

Appendices

Appendix A

Publications

The following publications have been produced in association with this thesis:

- F. Navarro, M. Escudero-Viñolo, J. Bescós: "SP-SIFT: Enhancing SIFT discrimination via super-pixel based foreground-background segregation." *IET Electronics Letters*, 2014, vol. 50, no 4, p. 272- 274, (Digital Object Identifier: 10.1049/el.2013.3949). Impact facrtor: 1,343 - Q2.
  - Chapter 4.
- F. Navarro, M. Escudero-Viñolo, Jesús Bescós: "Enhancing region-based object tracking with the SP-SIFT feature.", 2014, International Workshop on Content-Based Multimedia Indexing (pp. 1-4). IEEE.
  - Chapter 5.
- F. Navarro, M. Escudero-Viñolo, Jesús Bescós: "HPSTr: Homography Point-based Shape-fitted Tracker." *under review-2019*
  - Chapter 5
- F. Navarro, M. Escudero-Viñolo, Jesús Bescós: "Accurate segmentation and registration of skin lesion images to evaluate lesion change." IEEE Journal of Biomedical and Health Informatics, 2018, vol. 23, no 2, p. 501-508. (Digital Object Identifier: 10.1109/JBHI.2018.2825251). Impact facrtor: 4,217 - Q1.
  - Chapter 6
  - Chapter 7

The following reports or master thesis have been produced in association with this thesis:

- F. Navarro: "Local features and superpixels", 2014 Tech Report, Queen Mary University of London.
  - Chapter 2.
- M. Martín, F. Navarro: "Evaluación comparativa de técnicas de detección y descripción de putnos de interés en imágenes". Master Thesis, 2016, Univesidad Autónoma de Madrid
  - Chapter 2.
- F. Navarro, E. Velasco, J. Bescós: "Seguimiento basado en modelado dual RGB-D", 2016 National Conference on International Union of Radio Science (pp. 1-4).
  - Chapter 5.
- F. Navarro: "Viability study of video applications for outdoor cameras - UAV vision", 2016 Tech Report, Ref: UAM/45.
  - Chapter 2.

Appendix B

Logros, conclusiones y trabajo futuro

B.1 Resumen de logros y conclusiones principales

En esta tesis hemos consideramos el problema de los detectores/descriptores de características locales y los algoritmos de segmentación de regiones para la caracterización de imágenes, con el objetivo de encontrar una manera efectiva de combinarlos para mejorar sus capacidades individuales.

En primer lugar, hemos revisado ambas técnicas y destacado sus propiedades principales a considerar al desarrollar nuevos métodos y analizar los existentes (Capítulo 2). A continuación, hemos propuesto (Capítulo 3) una hipótesis para combinar sus capacidades que ha dado lugar a un par de estrategias de combinación, SP-SIFT y LF-SLIC, que conducen a una generalización efectiva.

La primera estrategia, SP-SIFT, se centra en la mejora de las características locales, concretamente en la característica local SIFT [Lowe, 2004], mediante el uso de algoritmos de segmentación de regiones—superpíxeles SLIC [Achanta et al., 2012]—para guiar el proceso de descripción de la información de la imagen (Capítulo 4).

La segunda, LF-SLIC, está enfocada en la mejora de las técnicas de segmentación de regiones, concretamente en los superpíxeles SLIC, usando detectores de características locales—detector de características locales SIFT—para guiar la gestión de la escala en el proceso de segmentación (Capítulo 6).

Para evaluar las hipotéticas ventajas de estas estrategias, hemos considerado dos aplicaciones diferentes. En primer lugar, hemos mostrado cómo usar la característica local SP-SIFT como una herramienta para mejorar las soluciones de seguimiento del estado del arte a través de su capacidad de descripción inhibiendo la información de fondo (Capítulo 5). Después, en el ámbito de la segmentación de regiones, hemos presentado una propuesta para usar el método de segmentación de regiones LF-SLIC en imagen médica, concretamente para segmentación de lesiones pigmentadas de la piel. Los resultados obtenidos sugieren que el esquema LF-SLIC se

beneficia de la inicialización de superpíxeles guiada por SIFT (Capítulo 7).

* * *

Los esquemas de combinación propuestos son ejemplos de las técnicas potentes que resultan de la combinación de características locales y algoritmos de segmentación de regiones, mejorando los métodos de referencia y generalmente mejorando el estado del arte para las aplicaciones de visión artificial estudiadas. Merece la pena estudiar los posibles campos de aplicación que pueden beneficiarse de las capacidades logradas. Destacar que el esquema de combinación no solo permite que los métodos trabajen juntos, sino que también tiene como objetivo que la combinación mejore sus fortalezas y minimice sus debilidades. En nuestra opinión, los futuros investigadores seguirán combinando características locales y algoritmos de segmentación de regiones con diferentes propiedades, con suerte obteniendo soluciones más potentes y versátiles adaptadas a sus aplicaciones objetivo.

B.2 Discusión de preguntas abiertas y trabajo futuro

Como suele ocurrir con muchos esfuerzos académicos, la investigación presentada en esta tesis abre muchas más preguntas que respuestas da. A continuación, discutimos sobre las principales partes de la tesis, algunas de las preguntas abiertas en ellas y las direcciones futuras que creemos se han de seguir para darles respuesta.

Parte II El objetivo de esta parte es presentar la hipótesis, y las dos afirmaciones que se evaluarán para validarla. Para ello, en el Capítulo 2 estudiamos el estado del arte tanto de las características locales como de los algoritmos de segmentación en regiones. Analizamos sus propiedades y elegimos dos técnicas que consideramos adecuadas para el desarrollo de la tesis: SIFT [Lowe, 2004] y SLIC [Achanta et al., 2012].

En el Capítulo 3, basado en las propiedades analizadas antes, presentamos la hipótesis de la tesis: **"La combinación de características locales—basados en descriptores locales—y algoritmos de segmentación en regiones—basados en relaciones espaciales y de color—dan lugar a una nueva familia de características de imagen con una mayor campo de aplicación gracias a sus fortalezas complementarias: capacidad discriminativa y robustez antes las variaciones en la información local"**. Dividimos la hipótesis en dos afirmaciones para poder validarla. En primer lugar, el uso de los algoritmos de segmentación en regiones es beneficioso para el proceso de descripción de las características locales. Segundo, el uso de los detectores de características locales es ventajoso para mejorar los resultados de la segmentación en regiones.

Para el trabajo futuro, podría ser interesante valorar la posibilidad de selección distintos métodos para las combinaciones. La idea es extender la estrategia propuesta a diferentes carac-

terísticas locales y algoritmos de segmentación en regiones. De hecho, SIFT y SLIC se eligieron, no solo por presentar resultados de primer nivel en sus respectivas categorías, sino también por ser considerados estándar. Técnicas similares podrían ser fácilmente integradas en el mismo esquema de combinación. Específicamente, proponemos dos combinaciones iniciales. En primer lugar, podría ser interesante ver cómo afecta la combinación a las características locales que usan patrones de descripción circulares. Sugerimos desarrollar SP-DAISY o SP-GLOH. En segundo lugar, podría ser de utilidad evaluar las características locales aprendidas, como TILDE o LIFT, para inicializar las técnicas de segmentación en regiones.

Parte III El objetivo de esta parte es definir una estrategia que combine con éxito las características locales y los algoritmos de segmentación en regiones de la manera que se propone en la primera afirmación. También es objetivo de esta parte el validar dicha propuesta en aplicaciones reales.

Con este fin, en el Capítulo 4 presentamos la característica local SP-SIFT. Un método que usa la segmentación en superpíxeles de SLIC para aislar la información que SIFT incluirá en su proceso de descripción. También hemos validado las características de este nuevo método. En primer lugar, hemos probado que conserva las capacidades de SIFT. Seguidamente, hemos probado que se han adquirido capacidades adicionales, como ser más robusto y discriminativo en escenarios con multitudes.

En el Capítulo 5 usamos SP-SIFT en una aplicación de visión artificial para verificar su capacidades teóricas. Hemos elegido seguimiento como la aplicación objetivo. Hemos propuesto una integración de la característica local en dos pasos. El objetivo es probar su rendimiento de manera secuencial, en primer lugar, sin que interfieran factores externos, y posteriormente utilizándolo como núcleo central para el seguimiento. Por lo tanto, inicialmente integramos SP-SIFT en un algoritmo de seguimiento del estado del arte para medir su potencial sin problemas adicionales derivados del método de seguimiento. A continuación, vamos un paso más allá y proponemos un algoritmo de seguimiento con SP-SIFT en el núcleo, definiendo estrategias innovadoras de seguimiento que solo son posibles gracias a las capacidades que aporta SP-SIFT.

El resultado de ambas aplicaciones confirma las capacidades de la técnica propuesta, así como validar la primera afirmación de la tesis en aplicaciones de seguimiento.

Relacionado con el trabajo futuro, sería interesante evaluar la propuesta en aplicaciones alternativas. Sugerimos evaluar su impacto en aplicaciones de imagen médica, un campo activo y con potencial de impacto en los últimos años.

Parte IV El objetivo de esta parte es definir una estrategia que combine con éxito las características locales y los algoritmos de segmentación en regiones de la manera que se propone en la segunda afirmación. Como en la Parte III, también es objetivo de esta parte validar dicha propuesta en aplicaciones reales.

En el Capítulo 6 presentamos el algoritmo de segmentación de regiones LF-SLIC. Un método que usa el detector de características locales de SIFT para mejorar los resultados de segmentación de los superpíxeles SLIC, mediante la obtención de superpíxeles a multi resolución sin tener que llevar a cabo ejecuciones de la segmentación a distintas escalas. En el mismo capítulo validamos las características del nuevo método. En primer lugar, comprobamos que los resultados en segmentación siguen siendo similares o mejores a los de SLIC. A continuación, comparamos las mejoras en eficiencia, esto es, el ahorro de coste computacional. Ambas evaluaciones resultan positivas, esto es, validan la afirmación, incluso pese a que se esperaba que fuera en la aplicación real donde se apreciara la mayor mejora de la propuesta.

En el Capítulo7 utilizamos LF-SLIC en una aplicación de visión artificial para validar sus capacidades. Hemos seleccionado la segmentación de lesiones en la piel como aplicación objetivo. Hemos propuesto una solución para aislar las lesiones de la piel de manera automática. La propuesta incluye LF-SLIC en el núcleo. Los resultados obtenidos validan tanto las capacidades de la nueva propuesta como la segunda afirmación de la tesis en la aplicación objetivo.

Como trabajo futuro, consideramos de interés evaluar el método propuesto en aplicaciones distintas. Sugerimos probarlo en aplicaciones de automoción, por ejemplo, para la detección de líneas y vehículos, un área de investigación recurrente cuyo avance, aunque constante, es más lento de lo esperado.

Finalmente, consideramos que ambas propuestas, su validación y los resultados de las aplicaciones, validan la hipótesis de la tesis en los esquemas probados.

Combinaciones alternativas De aquí en adelante, consideramos que podría ser interesante evaluar la combinación de técnicas distintas a las características locales y a los algoritmos de segmentación en regiones bajo el mismo esquema. Consideramos que habría únicamente dos restricciones con relación a que técnicas se podrían combinar. En primer lugar, para las combinaciones similares a SP-SIFT la única restricción sería combinar métodos que incluyan información de un entorno local en su descripción con métodos que integren algún nivel de agrupación de información de la imagen con contenido semántico. En segundo lugar, para las combinaciones similares a LF-SLIC la restricción sería usar técnicas que detecten puntos o áreas de interés en la imagen y, de nuevo, métodos que integren algún nivel de agrupación de información de la imagen con contenido semántico.

Específicamente, consideramos doy algoritmos de segmentación de regiones, mejorando los métodos de s propuestas iniciales. En primer lugar, probar técnicas de segmentación semántica [Kreso et al., 2017; Rota Bulò et al., 2018] para guiar el proceso de descripción de las características locales. La segmentación semántica se define como un proceso de segmentación de la imagen en partes que pertenecen a las mismas clases de objetos. En segundo lugar, podría resultar de utilidad evaluar modelos de atención visual [Zhao et al., 2015; Wang and Shen, 2017]

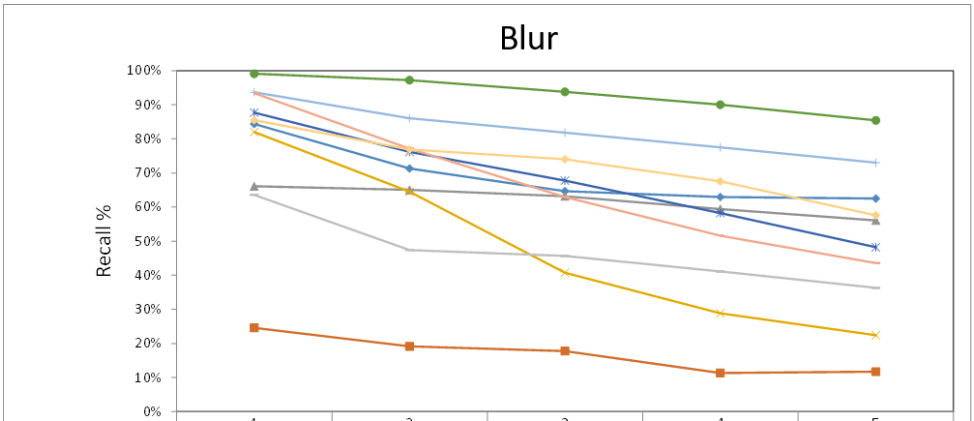
para inicializar técnicas de segmentación en regiones. Los métodos de atención visual intentan predecir la cantidad de atención que dirige el sistema visual y cognitivo humano a las distintas partes de una imagen.

Appendix C

LF-DT and LF-DS performance evaluation results

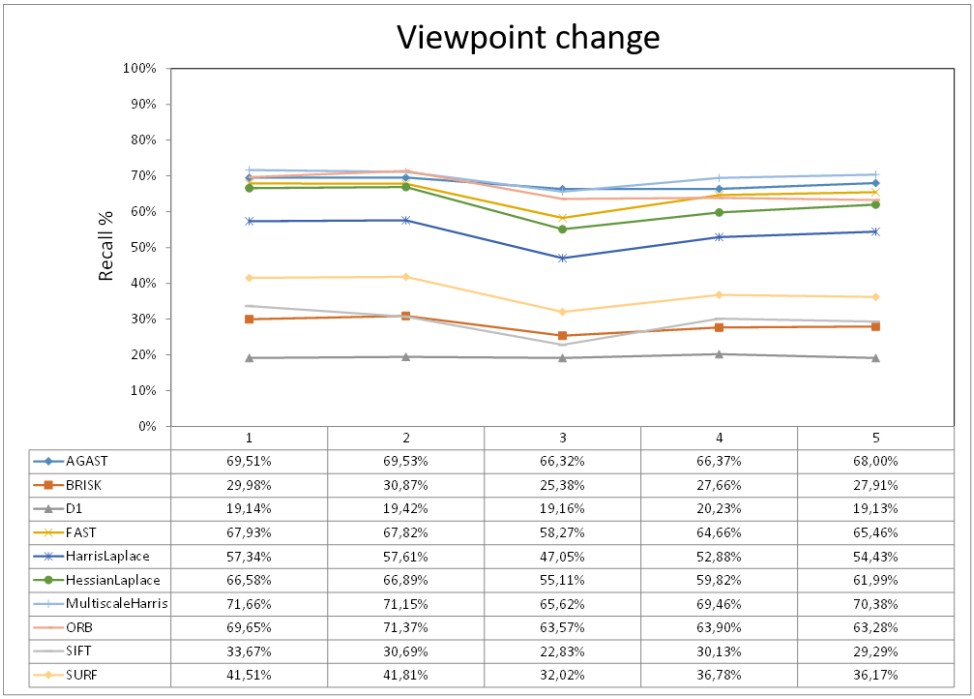
Results in terms of Recall and Precision are detailed per category. Figures and numbers are extracted from [Martín Redondo, 2016]

In the following figures we present the curves and numbers of the Recall (100xRecall) of the LF-DT for the different categories.



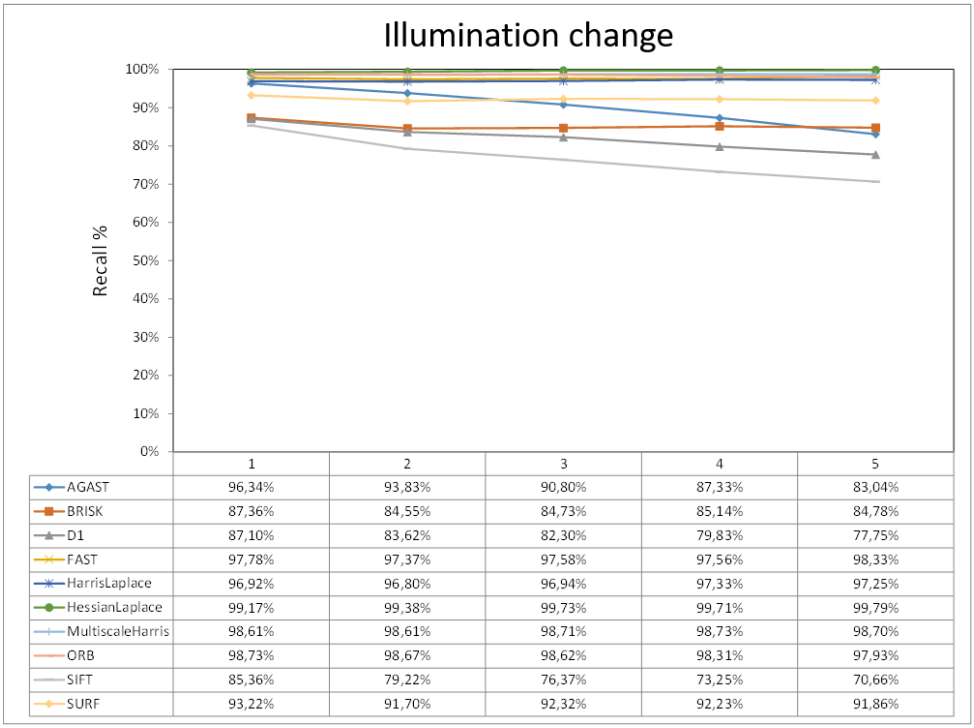
	1	2	3	4	5
AGAST	84,40%	71,30%	64,69%	62,96%	62,48%
BRISK	24,59%	19,13%	17,77%	11,34%	11,72%
D1	66,08%	65,05%	63,19%	59,41%	56,07%
FAST	81,97%	64,52%	40,74%	28,85%	22,44%
HarrisLaplace	87,69%	76,15%	67,74%	58,25%	48,24%
HessianLaplace	99,11%	97,19%	93,78%	90,04%	85,43%
MultiscaleHarris	93,66%	86,01%	81,81%	77,49%	73,02%
ORB	93,38%	77,12%	62,96%	51,58%	43,59%
SIFT	63,55%	47,38%	45,65%	41,14%	36,31%
SURF	85,50%	76,89%	74,01%	67,54%	57,56%

Fig. C.1. LF-DT Recall results for the isolated transformations (Blur) at global image.



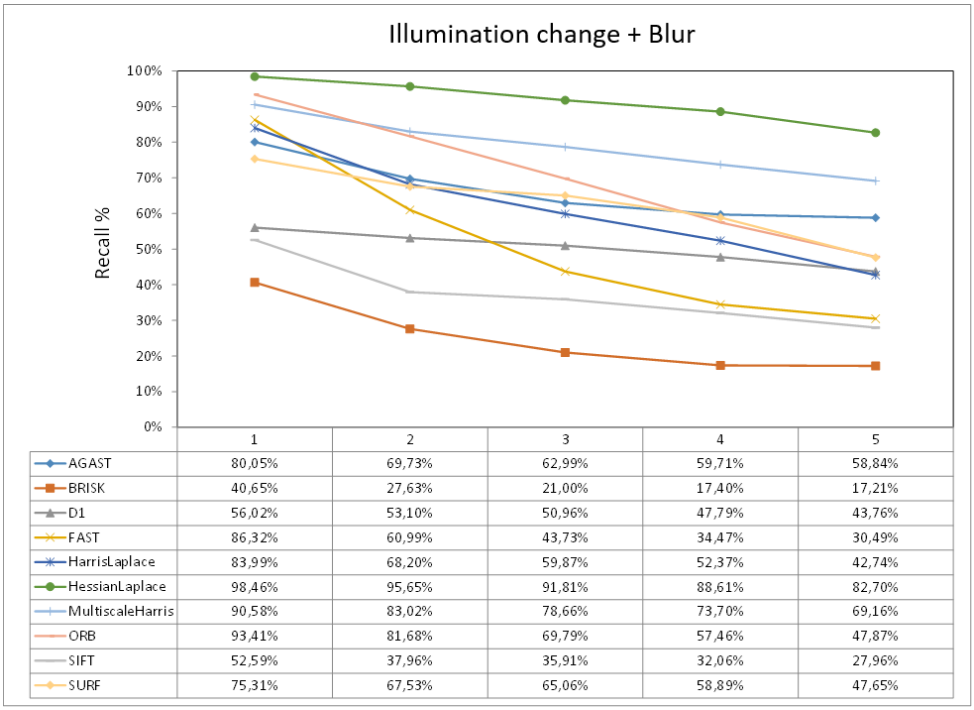
	1	2	3	4	5
AGAST	69,51%	69,53%	66,32%	66,37%	68,00%
BRISK	29,98%	30,87%	25,38%	27,66%	27,91%
D1	19,14%	19,42%	19,16%	20,23%	19,13%
FAST	67,93%	67,82%	58,27%	64,66%	65,46%
HarrisLaplace	57,34%	57,61%	47,05%	52,88%	54,43%
HessianLaplace	66,58%	66,89%	55,11%	59,82%	61,99%
MultiscaleHarris	71,66%	71,15%	65,62%	69,46%	70,38%
ORB	69,65%	71,37%	63,57%	63,90%	63,28%
SIFT	33,67%	30,69%	22,83%	30,13%	29,29%
SURF	41,51%	41,81%	32,02%	36,78%	36,17%

Fig. C.2. LF-DT Recall results for the isolated transformations (Viewpoint Change) at global image.



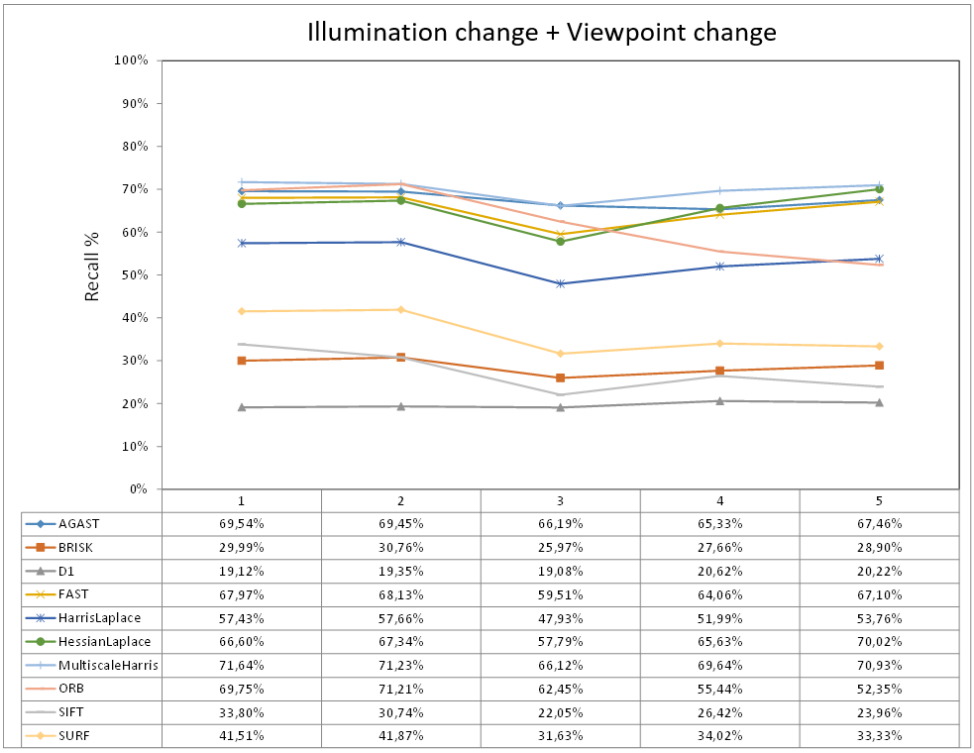
	1	2	3	4	5
AGAST	96,34%	93,83%	90,80%	87,33%	83,04%
BRISK	87,36%	84,55%	84,73%	85,14%	84,78%
D1	87,10%	83,62%	82,30%	79,83%	77,75%
FAST	97,78%	97,37%	97,58%	97,56%	98,33%
HarrisLaplace	96,92%	96,80%	96,94%	97,33%	97,25%
HessianLaplace	99,17%	99,38%	99,73%	99,71%	99,79%
MultiscaleHarris	98,61%	98,61%	98,71%	98,73%	98,70%
ORB	98,73%	98,67%	98,62%	98,31%	97,93%
SIFT	85,36%	79,22%	76,37%	73,25%	70,66%
SURF	93,22%	91,70%	92,32%	92,23%	91,86%

Fig. C.3. LF-DT Recall results for the isolated transformations (Illumination Change) at global image.



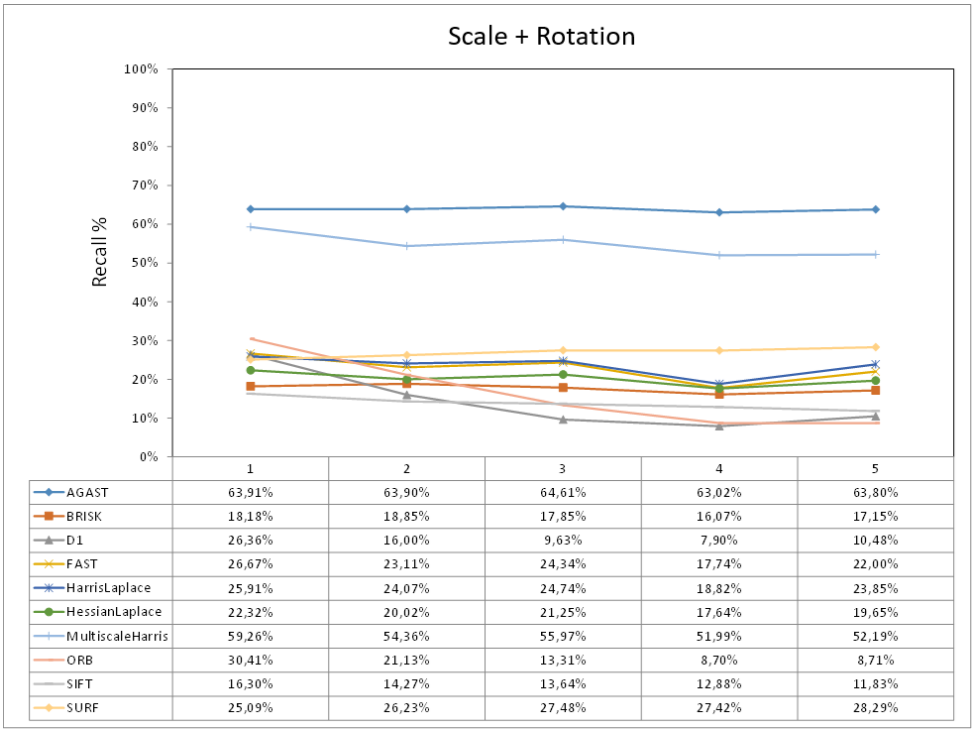
	1	2	3	4	5
AGAST	80,05%	69,73%	62,99%	59,71%	58,84%
BRISK	40,65%	27,63%	21,00%	17,40%	17,21%
D1	56,02%	53,10%	50,96%	47,79%	43,76%
FAST	86,32%	60,99%	43,73%	34,47%	30,49%
HarrisLaplace	83,99%	68,20%	59,87%	52,37%	42,74%
HessianLaplace	98,46%	95,65%	91,81%	88,61%	82,70%
MultiscaleHarris	90,58%	83,02%	78,66%	73,70%	69,16%
ORB	93,41%	81,68%	69,79%	57,46%	47,87%
SIFT	52,59%	37,96%	35,91%	32,06%	27,96%
SURF	75,31%	67,53%	65,06%	58,89%	47,65%

Fig. C.4. LF-DT Recall results for the combined transformations (Illumination Change and Blur) at global image.



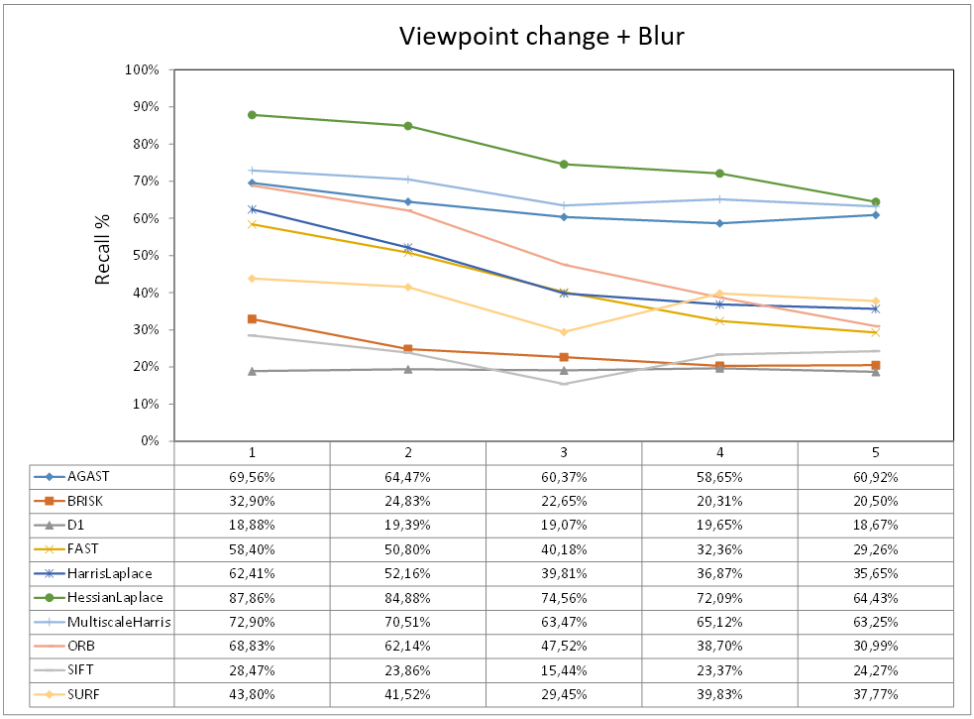
	1	2	3	4	5
AGAST	69,54%	69,45%	66,19%	65,33%	67,46%
BRISK	29,99%	30,76%	25,97%	27,66%	28,90%
D1	19,12%	19,35%	19,08%	20,62%	20,22%
FAST	67,97%	68,13%	59,51%	64,06%	67,10%
HarrisLaplace	57,43%	57,66%	47,93%	51,99%	53,76%
HessianLaplace	66,60%	67,34%	57,79%	65,63%	70,02%
MultiscaleHarris	71,64%	71,23%	66,12%	69,64%	70,93%
ORB	69,75%	71,21%	62,45%	55,44%	52,35%
SIFT	33,80%	30,74%	22,05%	26,42%	23,96%
SURF	41,51%	41,87%	31,63%	34,02%	33,33%

Fig. C.5. LF-DT Recall results for the combined transformations (Illumination Change and Viewpoint Change) at global image.



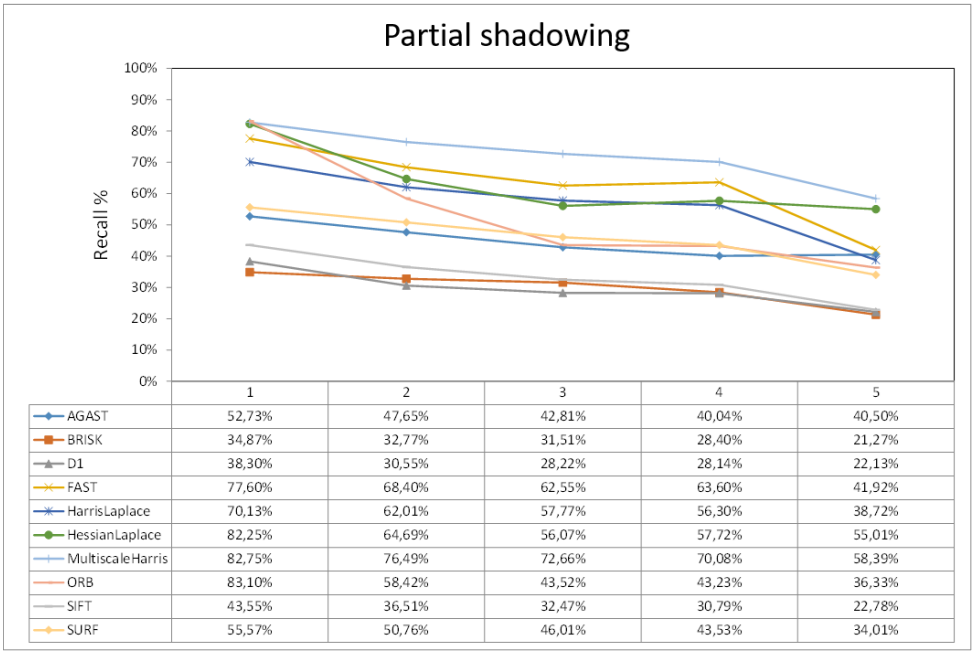
	1	2	3	4	5
AGAST	63,91%	63,90%	64,61%	63,02%	63,80%
BRISK	18,18%	18,85%	17,85%	16,07%	17,15%
D1	26,36%	16,00%	9,63%	7,90%	10,48%
FAST	26,67%	23,11%	24,34%	17,74%	22,00%
HarrisLaplace	25,91%	24,07%	24,74%	18,82%	23,85%
HessianLaplace	22,32%	20,02%	21,25%	17,64%	19,65%
MultiscaleHarris	59,26%	54,36%	55,97%	51,99%	52,19%
ORB	30,41%	21,13%	13,31%	8,70%	8,71%
SIFT	16,30%	14,27%	13,64%	12,88%	11,83%
SURF	25,09%	26,23%	27,48%	27,42%	28,29%

Fig. C.6. LF-DT Recall results for the combined transformations (Scale and Rotation) at global image.



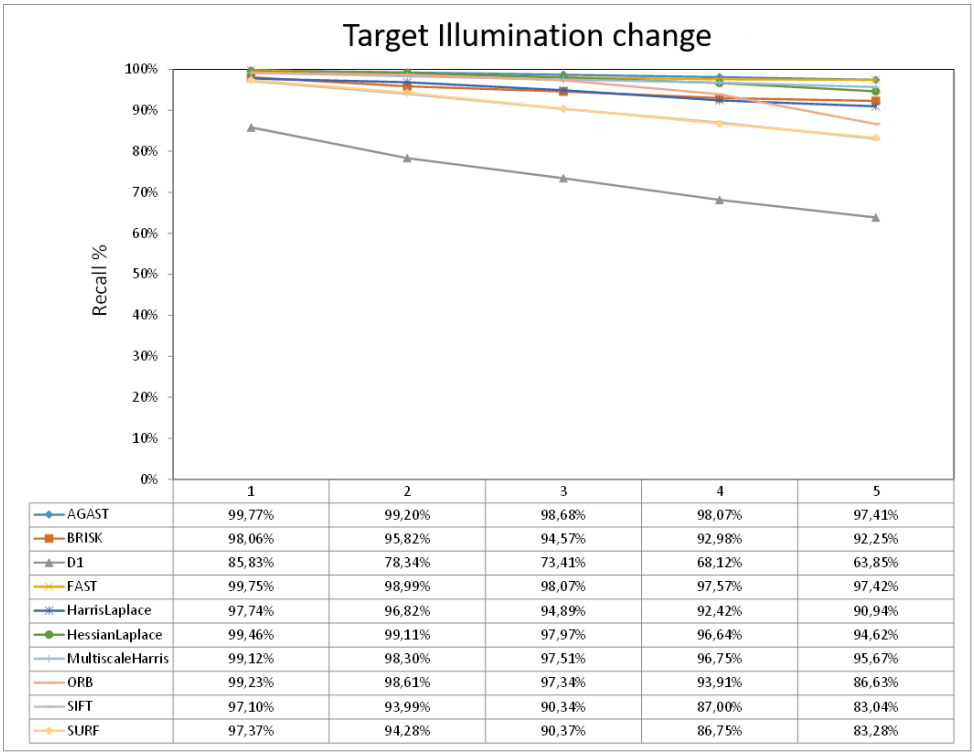
	1	2	3	4	5
AGAST	69,56%	64,47%	60,37%	58,65%	60,92%
BRISK	32,90%	24,83%	22,65%	20,31%	20,50%
D1	18,88%	19,39%	19,07%	19,65%	18,67%
FAST	58,40%	50,80%	40,18%	32,36%	29,26%
HarrisLaplace	62,41%	52,16%	39,81%	36,87%	35,65%
HessianLaplace	87,86%	84,88%	74,56%	72,09%	64,43%
MultiscaleHarris	72,90%	70,51%	63,47%	65,12%	63,25%
ORB	68,83%	62,14%	47,52%	38,70%	30,99%
SIFT	28,47%	23,86%	15,44%	23,37%	24,27%
SURF	43,80%	41,52%	29,45%	39,83%	37,77%

Fig. C.7. LF-DT Recall results for the combined transformations (Viewpoint Change and Blur) at global image.



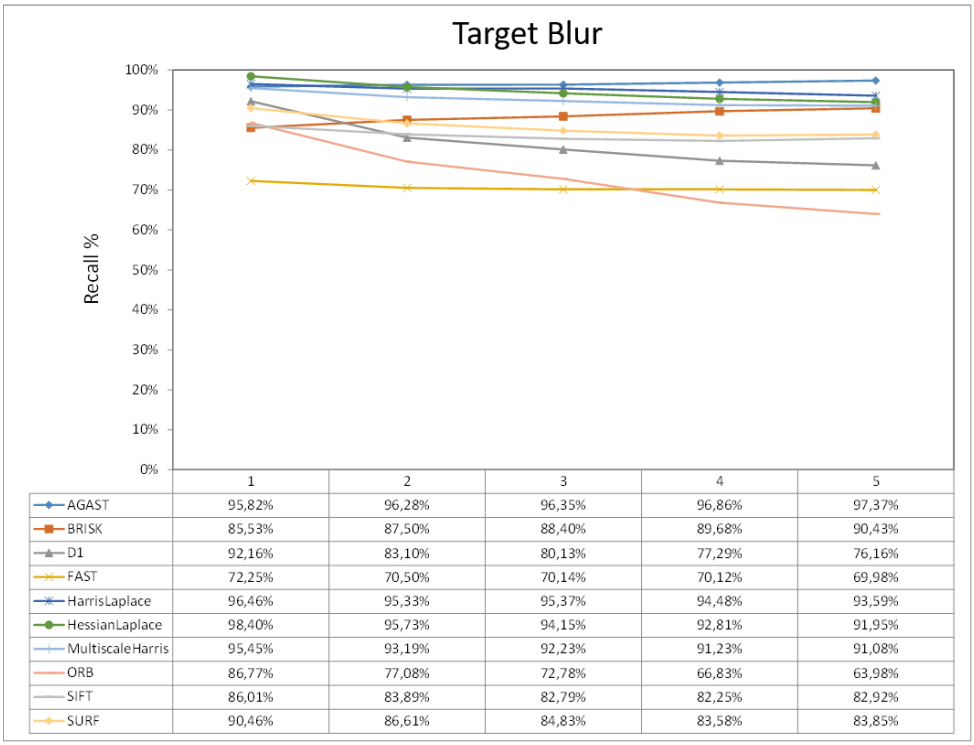
	1	2	3	4	5
AGAST	52,73%	47,65%	42,81%	40,04%	40,50%
BRISK	34,87%	32,77%	31,51%	28,40%	21,27%
D1	38,30%	30,55%	28,22%	28,14%	22,13%
FAST	77,60%	68,40%	62,55%	63,60%	41,92%
HarrisLaplace	70,13%	62,01%	57,77%	56,30%	38,72%
HessianLaplace	82,25%	64,69%	56,07%	57,72%	55,01%
MultiscaleHarris	82,75%	76,49%	72,66%	70,08%	58,39%
ORB	83,10%	58,42%	43,52%	43,23%	36,33%
SIFT	43,55%	36,51%	32,47%	30,79%	22,78%
SURF	55,57%	50,76%	46,01%	43,53%	34,01%

Fig. C.8. LF-DT Recall results for the isolated transformations (Partial Shadowing) at target level.



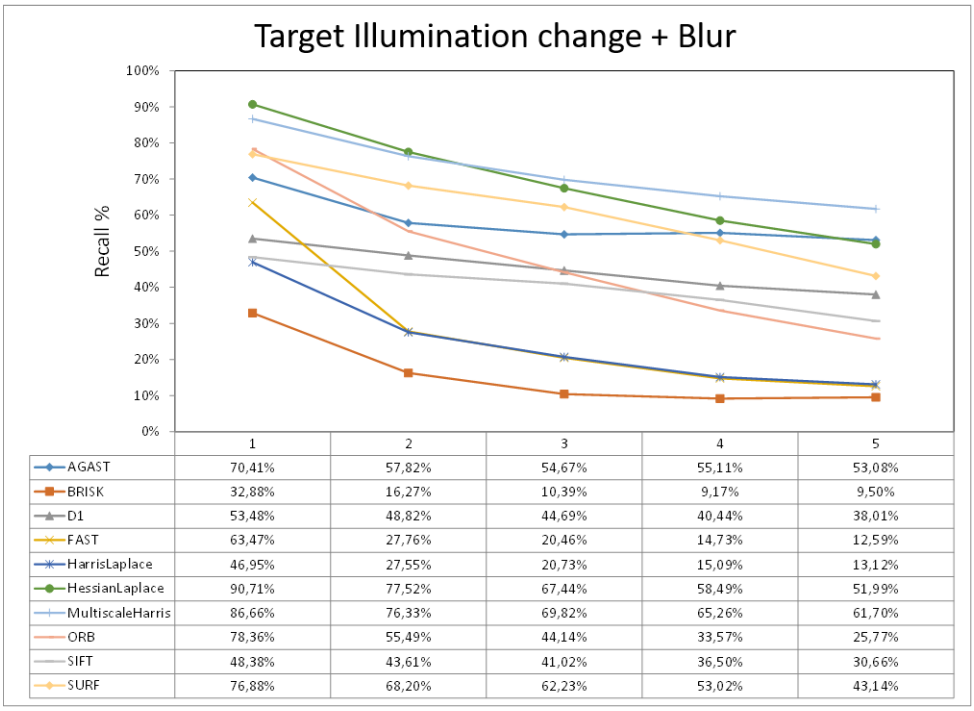
	1	2	3	4	5
AGAST	99,77%	99,20%	98,68%	98,07%	97,41%
BRISK	98,06%	95,82%	94,57%	92,98%	92,25%
D1	85,83%	78,34%	73,41%	68,12%	63,85%
FAST	99,75%	98,99%	98,07%	97,57%	97,42%
HarrisLaplace	97,74%	96,82%	94,89%	92,42%	90,94%
HessianLaplace	99,46%	99,11%	97,97%	96,64%	94,62%
MultiscaleHarris	99,12%	98,30%	97,51%	96,75%	95,67%
ORB	99,23%	98,61%	97,34%	93,91%	86,63%
SIFT	97,10%	93,99%	90,34%	87,00%	83,04%
SURF	97,37%	94,28%	90,37%	86,75%	83,28%

Fig. C.9. LF-DT Recall results for the isolated transformations (Illumination Change) at target level.



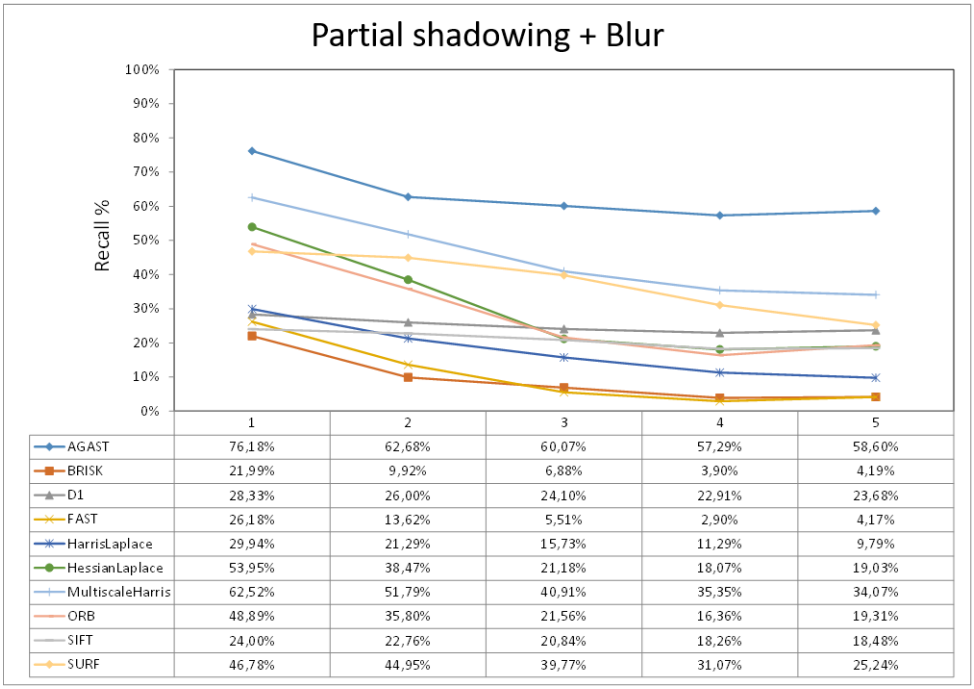
	1	2	3	4	5
AGAST	95,82%	96,28%	96,35%	96,86%	97,37%
BRISK	85,53%	87,50%	88,40%	89,68%	90,43%
D1	92,16%	83,10%	80,13%	77,29%	76,16%
FAST	72,25%	70,50%	70,14%	70,12%	69,98%
HarrisLaplace	96,46%	95,33%	95,37%	94,48%	93,59%
HessianLaplace	98,40%	95,73%	94,15%	92,81%	91,95%
MultiscaleHarris	95,45%	93,19%	92,23%	91,23%	91,08%
ORB	86,77%	77,08%	72,78%	66,83%	63,98%
SIFT	86,01%	83,89%	82,79%	82,25%	82,92%
SURF	90,46%	86,61%	84,83%	83,58%	83,85%

Fig. C.10. LF-DT Recall results for the isolated transformations (Blur) at target level.



	1	2	3	4	5
AGAST	70,41%	57,82%	54,67%	55,11%	53,08%
BRISK	32,88%	16,27%	10,39%	9,17%	9,50%
D1	53,48%	48,82%	44,69%	40,44%	38,01%
FAST	63,47%	27,76%	20,46%	14,73%	12,59%
HarrisLaplace	46,95%	27,55%	20,73%	15,09%	13,12%
HessianLaplace	90,71%	77,52%	67,44%	58,49%	51,99%
MultiscaleHarris	86,66%	76,33%	69,82%	65,26%	61,70%
ORB	78,36%	55,49%	44,14%	33,57%	25,77%
SIFT	48,38%	43,61%	41,02%	36,50%	30,66%
SURF	76,88%	68,20%	62,23%	53,02%	43,14%

Fig. C.11. LF-DT Recall results for the combined transformations (Illumination Change and Blur) at target level.



	1	2	3	4	5
AGAST	76,18%	62,68%	60,07%	57,29%	58,60%
BRISK	21,99%	9,92%	6,88%	3,90%	4,19%
D1	28,33%	26,00%	24,10%	22,91%	23,68%
FAST	26,18%	13,62%	5,51%	2,90%	4,17%
HarrisLaplace	29,94%	21,29%	15,73%	11,29%	9,79%
HessianLaplace	53,95%	38,47%	21,18%	18,07%	19,03%
MultiscaleHarris	62,52%	51,79%	40,91%	35,35%	34,07%
ORB	48,89%	35,80%	21,56%	16,36%	19,31%
SIFT	24,00%	22,76%	20,84%	18,26%	18,48%
SURF	46,78%	44,95%	39,77%	31,07%	25,24%

Fig. C.12. LF-DT Recall results for the combined transformations (Partial Shadowing and Blur) at target level.

Second, in the following figures we present the curves of the 1-Precision of the LF-DS for the different categories.

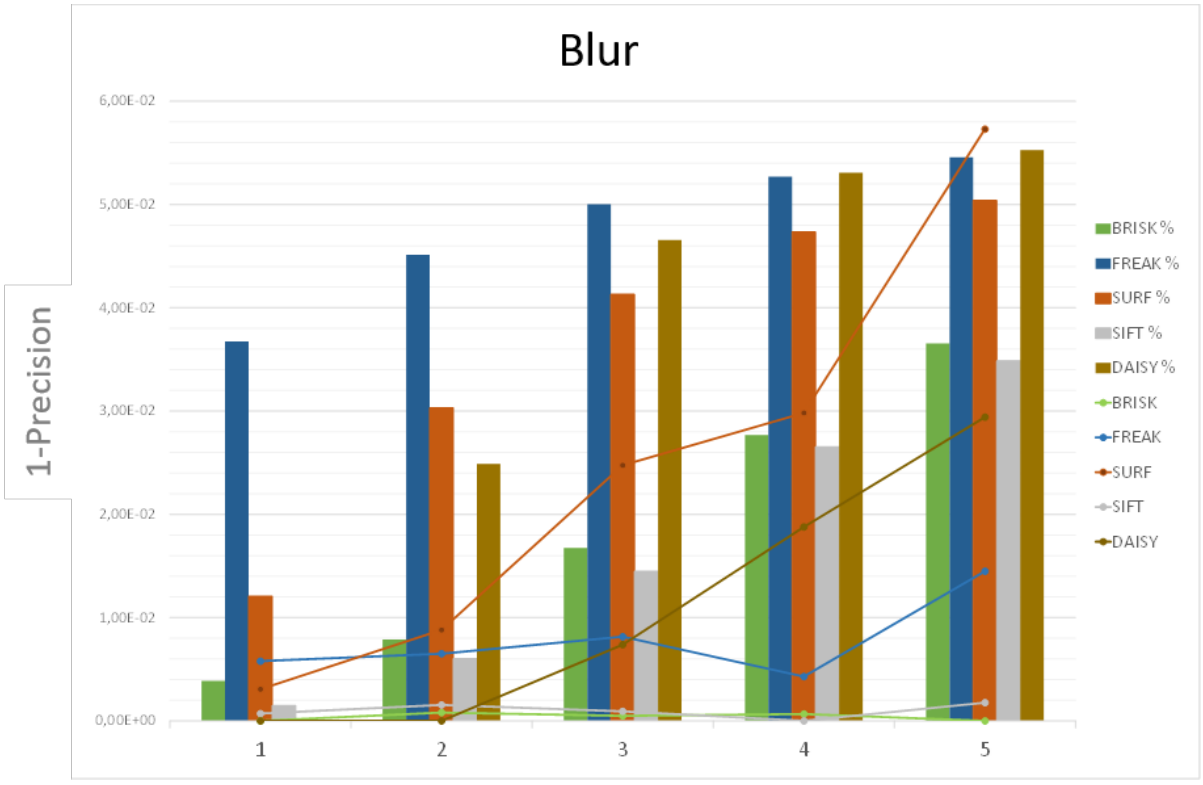


Fig. C.13. LF-DS Precision results for the isolated transformations (Blur) at global image.

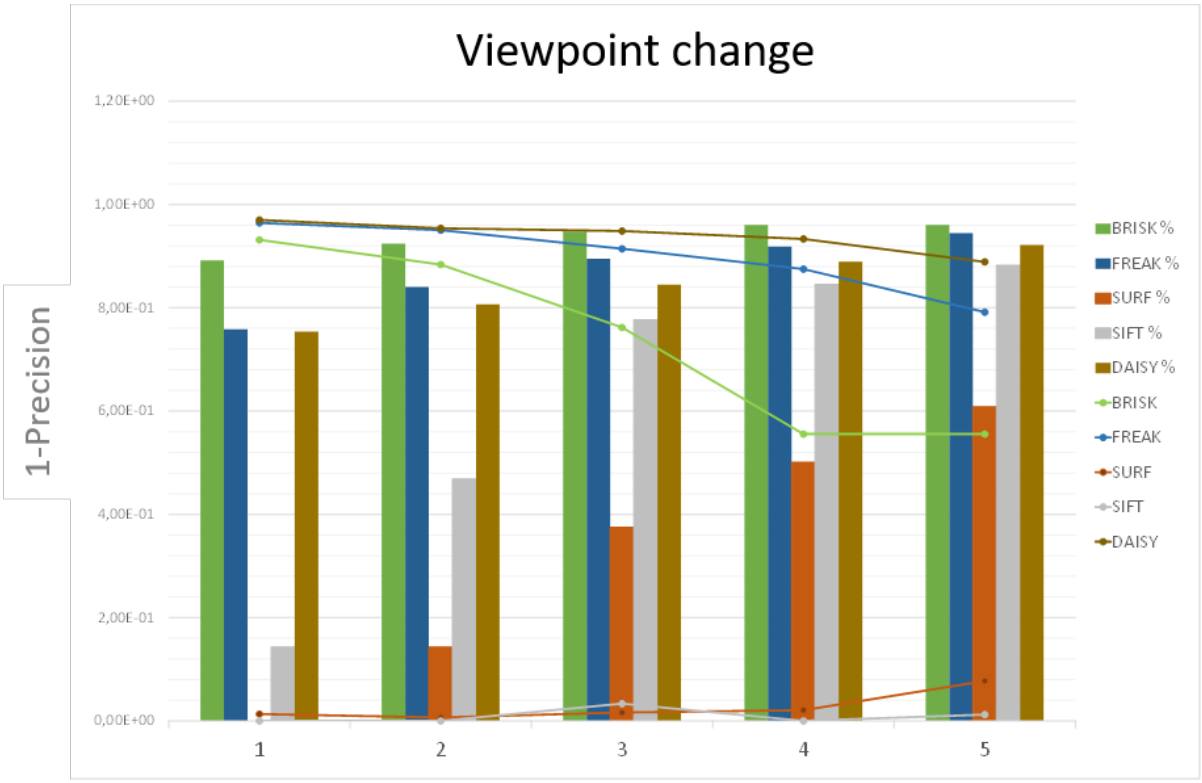


Fig. C.14. LF-DS Precision results for the isolated transformations (Viewpoint Change) at global image.

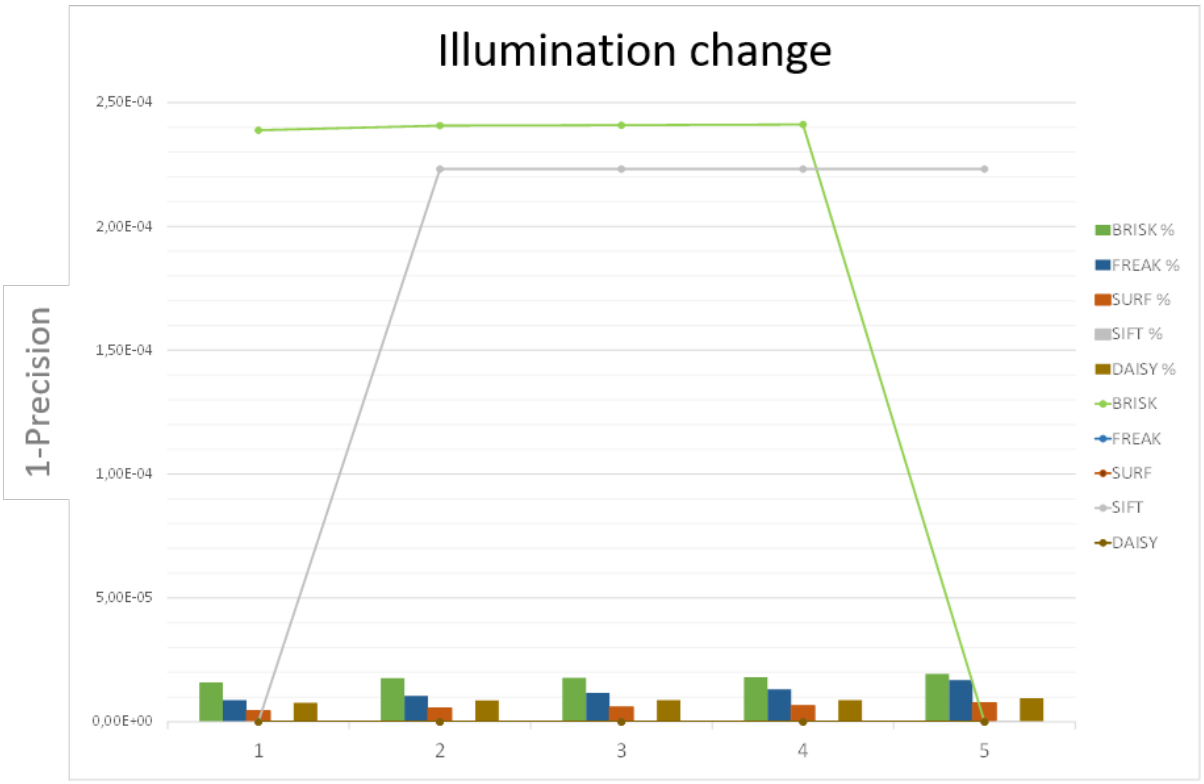


Fig. C.15. LF-DS Precision results for the isolated transformations (Illumination Change) at global image.

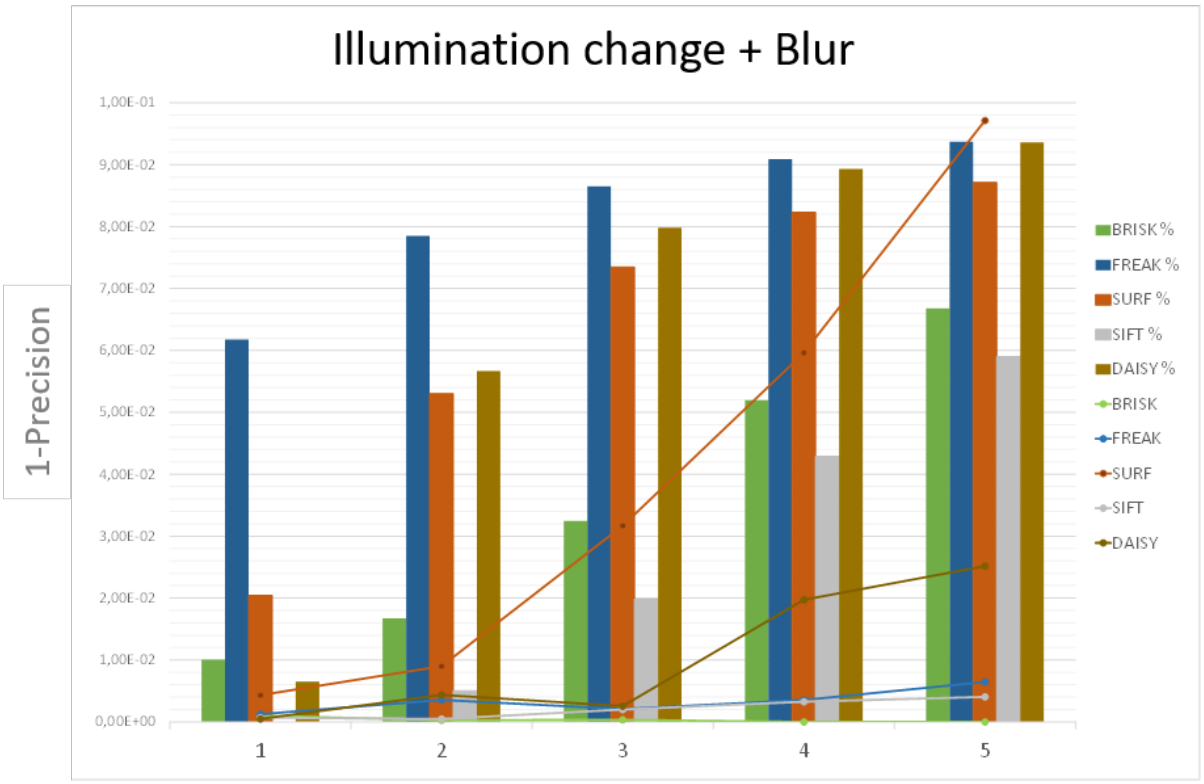


Fig. C.16. LF-DS Precision results for the combined transformations (Illumination Change and Blur) at global image.

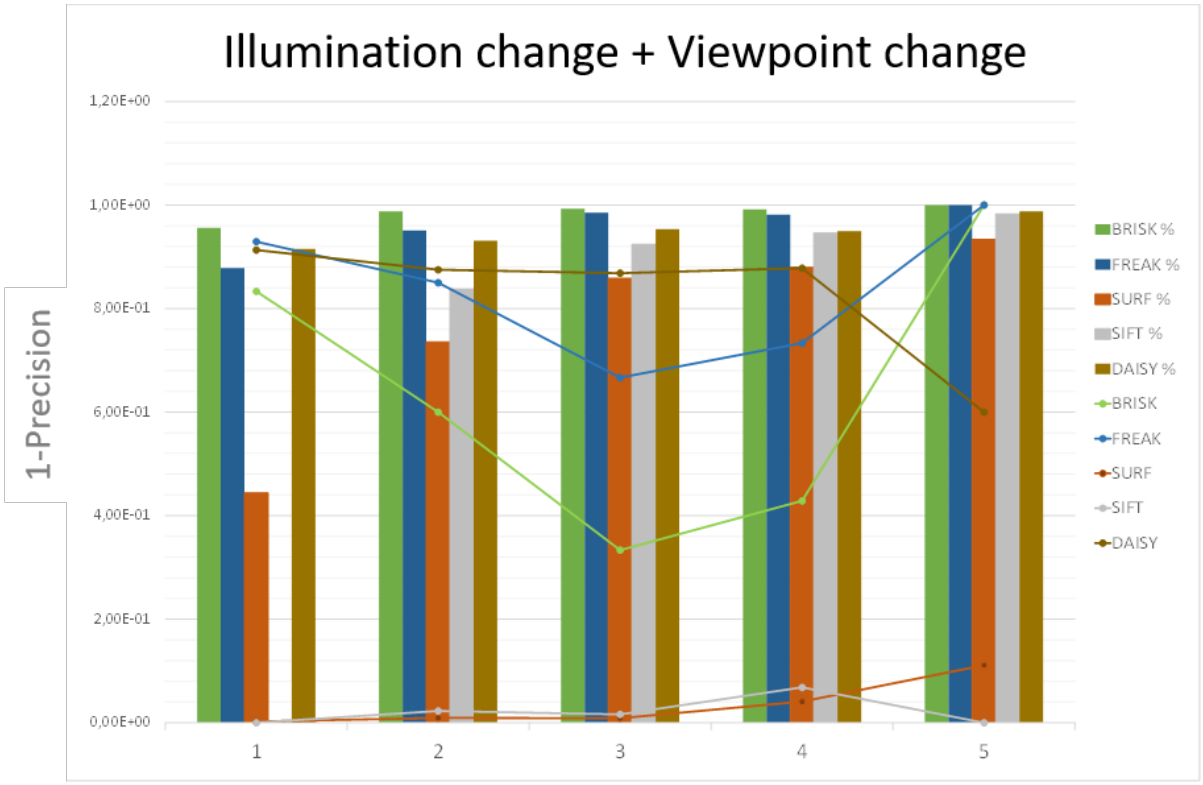


Fig. C.17. LF-DS Precision results for the combined transformations (Illumination Change and Viewpoint Change) at global image.

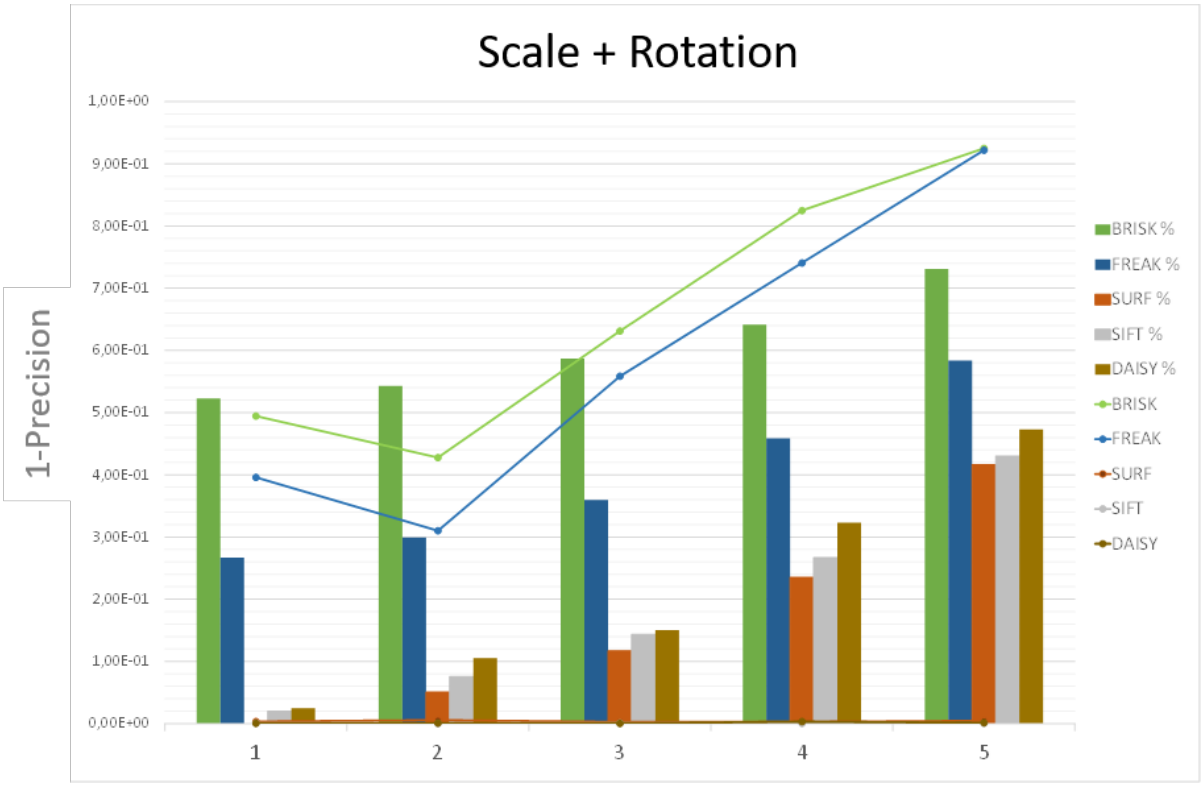


Fig. C.18. LF-DS Precision results for the combined transformations (Scale and Rotation) at global image.

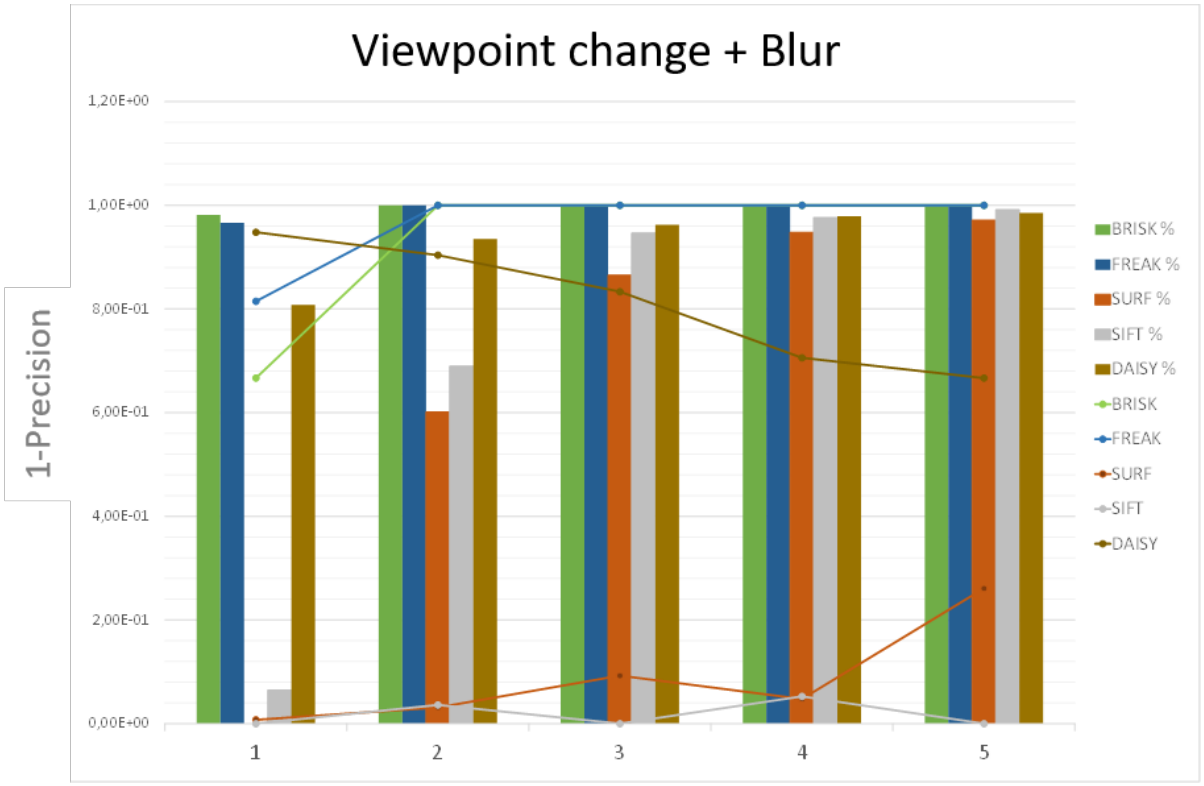


Fig. C.19. LF-DS Precision results for the combined transformations (Viewpoint Change and Blur) at global image.

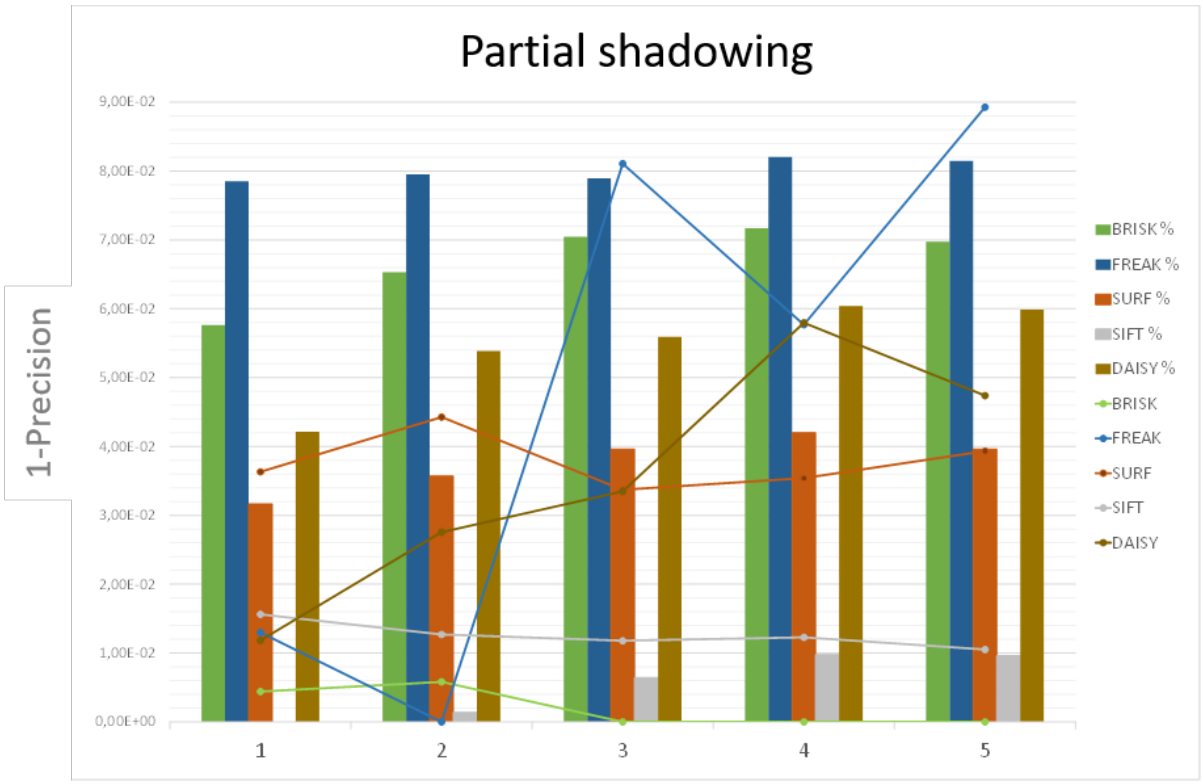


Fig. C.20. LF-DS Precision results for the isolated transformations (Partial Shadowing) at target level.

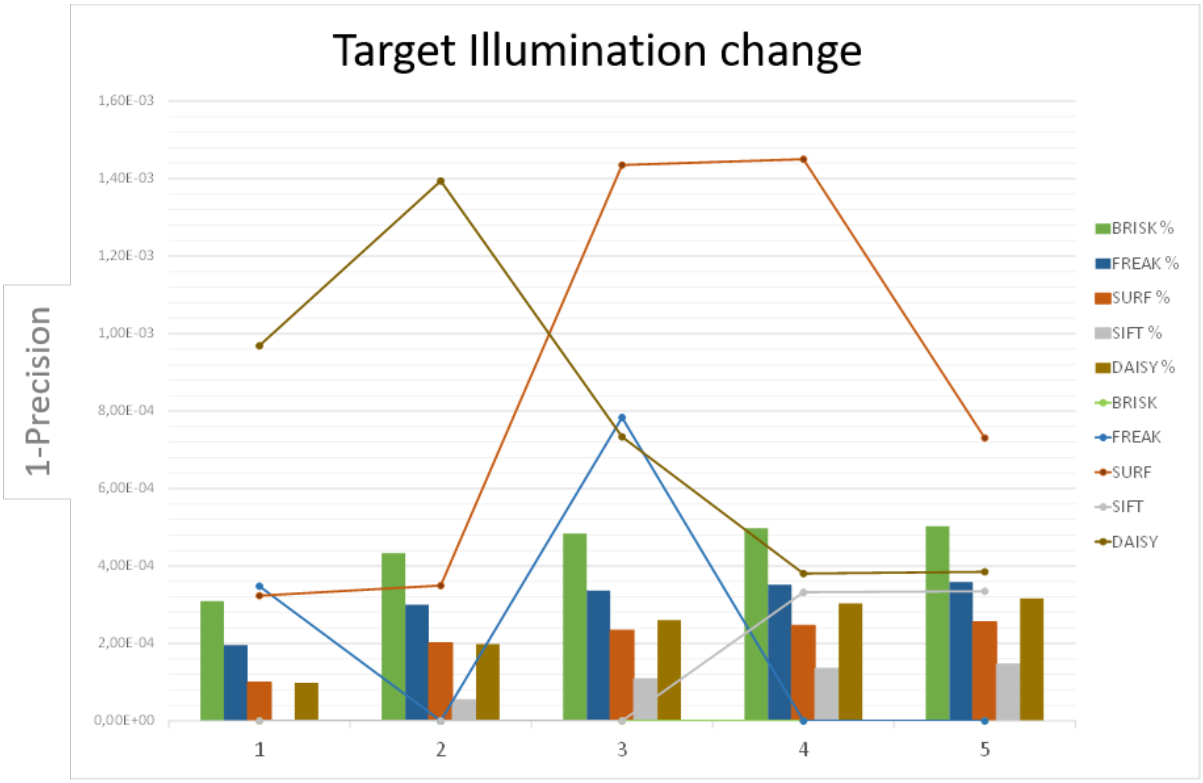


Fig. C.21. LF-DS Precision results for the isolated transformations (Illumination Change) at target level.

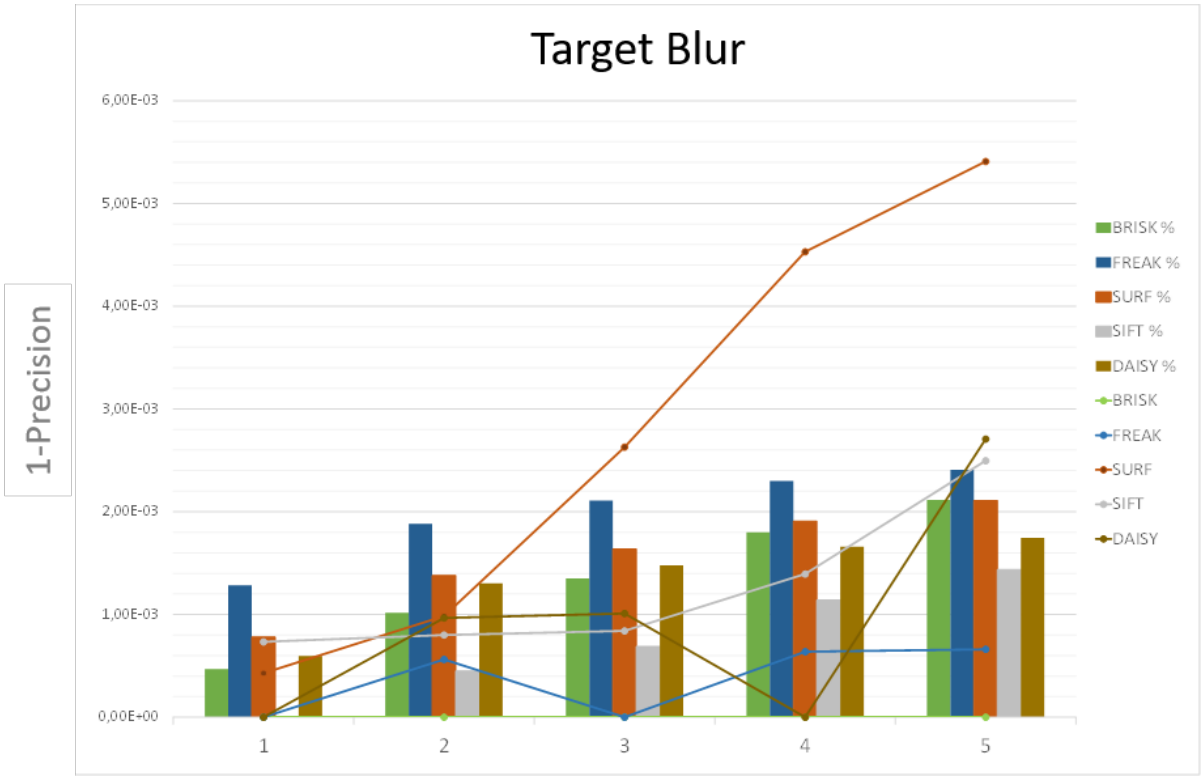


Fig. C.22. LF-DS Precision results for the isolated transformations (Blur) at target level.

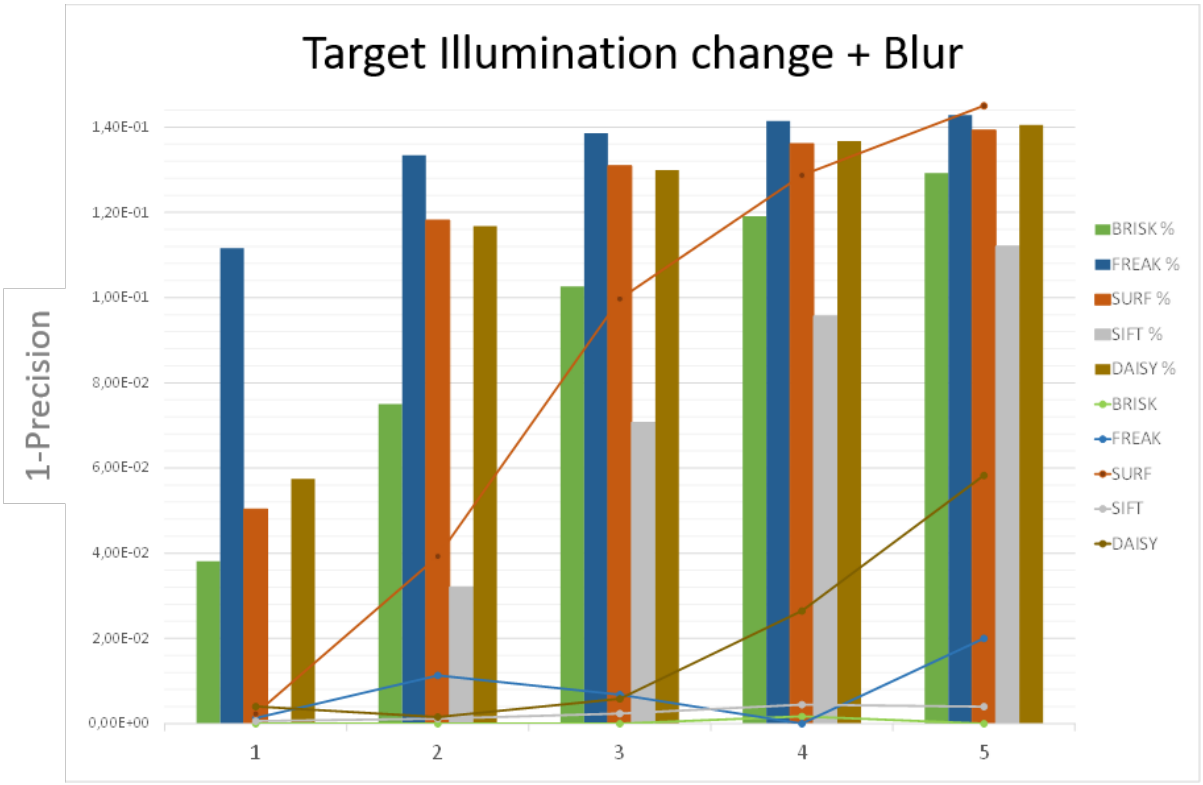


Fig. C.23. LF-DS Precision results for the combined transformations at target level.

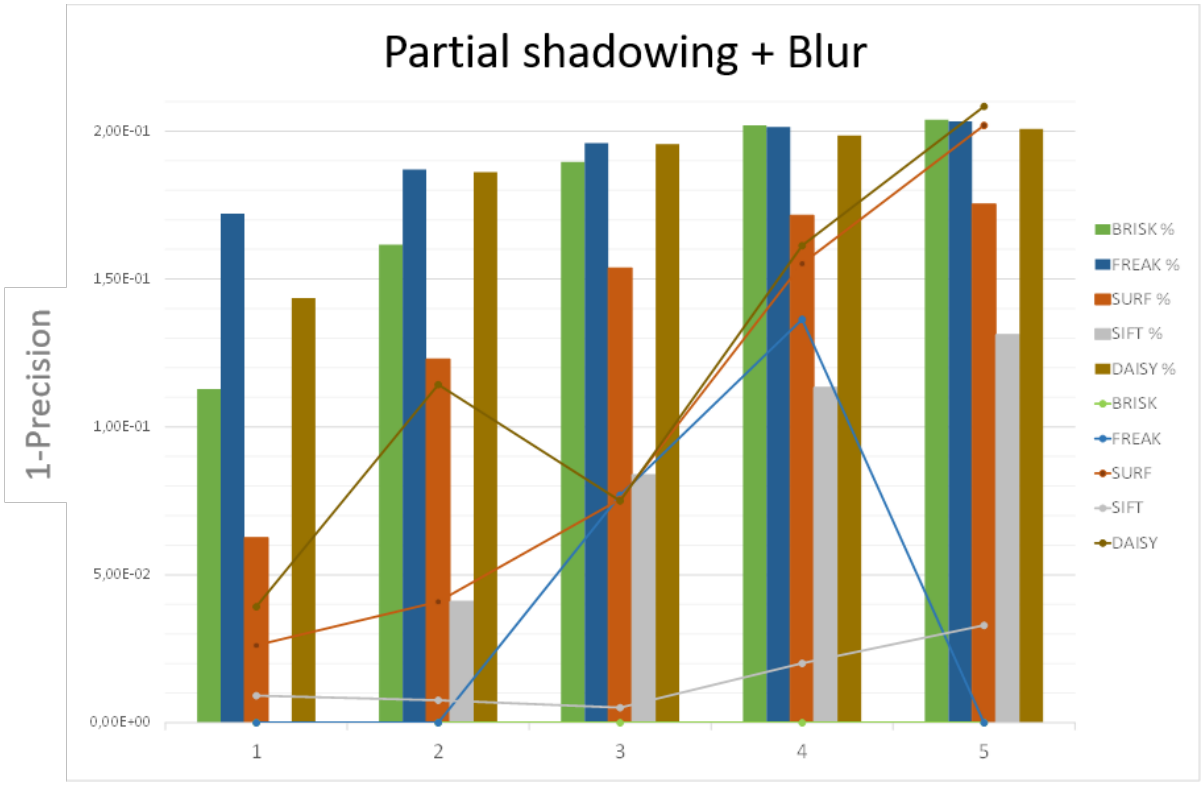


Fig. C.24. LF-DS Precision results for the combined transformations at target level.

Glossary

ACC *Accuracy*

AGAST *Adaptative and Generic Accelerated Segment Test*

AI *Artificial Intelligence*

BB *Bounding-Box*

BM *Binary Mask*

BPLR *Boundary Preserving Local Regions*

BRIEF *Binary Robust Independent Elementary Features*

BRISK *Binary Robust Invariant Scalable KeyPoint*

CAD *Computer-Aided Diagnosis*

CNNs *Convolutional Neural Networks*

ConvOpt *Convex Optimisation*

CSI *Composite Statistical Interference*

DeepDesc *Deep Descriptor*

DoG *Difference-of-Gaussians*

DynGraph *Dynamic Graph*

EBR *Edge-Based Regions*

EGraph *Efficient Hierarchical Graph-Based Video Segmentation*

ELM *Epiluminescence Microscopy*

ERS *Entropy Rate Superpixel*

EWCVT *Edge-Weighted Centroidal Voronoi Tessellations*

FLOG *Fan Laplacian of Gaussian*

Frag *Fragment-Based Tracking*

FREAK *Fast Retina KeyPoint*

GLOH *Gradiant Localization Oriented Histogram*

HB *Hough-Based Tracking*

HEWCVT *Hierarchical Edge-Weighted Centroidal Voronoi Tessellations*

HOG *Histograms of Oriented Gradients*

HPSTr *Homography Point-based Shape-fitted Tracker*

HSI *Hue, Saturation and Intensity*

HSV *Hue Saturation Value*

HT *Hough Transform*

IBR *Intensity Extrema-Based Region*

ISIC *International Skin Imaging Collaboration*

IVT *Incremental Learning for Visual Tracking*

JPEG *Joint Photographic Experts Group*

JOTS *Joint Online Tracking and Segmentation*

KeySeg *KeyPoint Segmentation*

KNN *K-nearest neighbors algorithm*

LBP *Local Binary Patterns*

LF *Local Features*

LF-DS *Local Features Description*

LF-DT *Local Features Detection*

LF-SLIC *Local Features Simple Linear Iterative Clustering*

LIFT *Learned Invariant Feature Transform*

LSC *Linear Spectral Clustering*

MAP *Maximum a Posteriori*

MFD *Medial Feature Detector*

MIL *Multiple Instance Learning*

MRFs *Markov Random Fields*

MS *Mean Shift*

MSER *Maximally Stable Extremal Region*

MSPDG *Multiscale Symmetric Part Detection and Grouping*

N-cut *Normalized cut*

OLT *Online Learning Trackers*

ORB *Oriented Fast and Rotated Brief*

OTS *Object Tracking by Segmentation*

PB *Pseudo-Boolean*

PCA *Principal Component Analysis*

PCBR *Principal Curvature-Based Region*

PDAT *Patch-Based Dynamic Appearance Tracking*

PETS *Performance Evaluation of Tracking and Surveillance*

PF *Particle Filter*

PhD *Doctor of Philosophy*

POIs *Points of Interests*

PROST *Parallel Robust Online Simple Tracking*

R-SPTrack *Robust Superpixels Tracking*

RGB *Red Green Blue*

ROI *Region of Interest*

SEEDS *Superpixels Extracted Via Energy-Driven Sampling*

SIFT *Scale Invariant Feature Transform*

SIFT-LBP *Scale Invariant Feature Transform - Local Binary Patterns*

SLIC *Simple Linear Iterative Clustering*

SOtA *State-Of-the-Art*

SP-SIFT *Superpixels Scale Invariant Feature Transform*

SPT *Superpixels tracking*

SPTrack *Superpixels Tracking*

Struck *Structured Output Tracking*

SURF *Speeded Up Robust Features*

TFeat *Triplets Features*

TILDE *Temporally Invariant Learned Detector*

TLD *Tracking Learning Detection*

VOS *Video Object Segmentation*

VTD *Visual Tracking Decomposition*

Bibliography

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 34(11):2274–2282, 2012. [Cited on pages 24, 25, 28, 29, 33, 42, 68, 70, 87, 88, 93, 117, 118, 125, and 126.]

A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *Computer Vision and Pattern Recognition (CVPR), 2006 IEEE Conference on*, volume 1, pages 798–805. IEEE, 2006. [Cited on pages 60, 66, and 79.]

A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 510–517. Ieee, 2012. [Cited on pages 18 and 22.]

P. F. Alcantarilla, A. Bartoli, and A. J. Davison. Kaze features. In *European Conference on Computer Vision*, pages 214–227. Springer, 2012. [Cited on page 21.]

C. N. E. Anagnostopoulos, D. D. Vergados, and P. Mintzias. Image registration of follow-up examinations in digital dermoscopy. In *BioInformatics and BioEngineering, 2013 IEEE International Conference on*, pages 1–4. IEEE, 2013. [Cited on page 102.]

Y. Avrithis and K. Rapantzikos. The medial feature detector: Stable regions from image boundaries. In *International Conference on Computer Vision*, pages 1724–1731. IEEE, 2011. [Cited on page 44.]

A. Ayvaci and S. Soatto. Motion segmentation with occlusions on the superpixel graph. In *ICCV Workshops*, pages 727–734, 2009. [Cited on page 25.]

B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*, pages 983–990. IEEE, 2009. [Cited on pages 63, 66, and 80.]

S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004. [Cited on pages 56 and 57.]

V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *British Machine Vision Conference*, volume 1, page 3, 2016. [Cited on pages 18 and 23.]

C. A. Z. Barcelos and V. Pires. An automatic based nonlinear diffusion equations scheme for skin lesion segmentation. *Applied Mathematics and Computation*, 215(1):251–261, 2009. [Cited on page 101.]

H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006*, pages 404–417. Springer, 2006. [Cited on pages 17, 18, 20, 41, and 108.]

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. [Cited on pages 5 and 71.]

P. R. Beaudet. Rotationally invariant image operators. In *Proc. 4th Int. Joint Conf. Pattern Recog, Tokyo, Japan, 1978*, 1978. [Cited on page 19.]

B. E. Bejnordi, G. Litjens, M. Hermsen, N. Karssemeijer, and J. A. van der Laak. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In *Medical Imaging 2015: Digital Pathology*, volume 9420, page 94200H. International Society for Optics and Photonics, 2015. [Cited on page 90.]

A. Ben-David and E. Frank. Accuracy of machine learning models versus hand crafted expert systems–a credit scoring case study. *Expert Systems with Applications*, 36(3):5264–5271, 2009. [Cited on page 3.]

M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74(1):59–73, 2007. [Cited on page 71.]

Z. Cai, L. Wen, Z. Lei, N. Vasconcelos, and S. Z. Li. Robust deformable and occluded object tracking with dynamic graph. *Image Processing, IEEE Transactions on*, 23(12):5497–5509, 2014. [Cited on pages 66 and 77.]

M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In *In proc. European Conference on Computer Vision*, pages 778–792. Springer, 2010. [Cited on pages 17, 18, and 21.]

L. Čehovin, A. Leonardis, and M. Kristan. Visual object tracking performance measures revisited. *arXiv preprint arXiv:1502.05803*, 2015. [Cited on page 63.]

M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V. Stoecker. Lesion border detection in dermoscopy images. *Computerized medical imaging and graphics*, 33(2):148–153, 2009. [Cited on page 101.]

D. H. Chung and G. Sapiro. Segmenting skin lesions with partial-differential-equations-based image processing algorithms. *IEEE transactions on Medical Imaging*, 19(7):763–767, 2000. [Cited on page 101.]

N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *Symposium on Biomedical Imaging (ISBI), 2018 IEEE*, pages 168–172. IEEE, 2018. [Cited on pages xii and 108.]

R. T. Collins. Mean-shift blob tracking through scale space. In *Computer Vision and Pattern Recognition (CVPR), 2003 IEEE Conference on*, volume 2, pages II–234. IEEE, 2003. [Cited on pages 66 and 79.]

R. T. Collins, Y. Liu, and M. Leordeanu. Online selection of discriminative tracking features. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1631–1643, 2005. [Cited on page 63.]

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):603–619, 2002. [Cited on page 104.]

D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000. [Cited on pages 56 and 57.]

C. Cui and K. N. Ngan. Scale-and affine-invariant fan feature. *IEEE Transactions on Image Processing*, 20(6):1627–1640, 2011. [Cited on page 44.]

H. Deng, W. Zhang, E. Mortensen, T. Dietterich, and L. Shapiro. Principal curvature-based region detector for object recognition. In *Computer Vision and Pattern Recognition (CVPR), 2007 IEEE Conference on*, pages 1–8. IEEE, 2007. [Cited on page 44.]

P. Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). https://github.com/pdollar/toolbox. [Cited on page 68.]

M. Donoser and H. Bischof. Efficient maximally stable extremal region (mser) tracking. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 553–560. Ieee, 2006. [Cited on page 44.]

S. Duffner and C. Garcia. Pixeltrack: a fast adaptive algorithm for tracking non-rigid objects. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2480–2487, 2013. [Cited on page 63.]

M. Escudero-Viñolo. *Contributions to region-based image and video analysis: feature aggregation, background subtraction and description constraining*. PhD thesis, Universidad Autónoma de Madrid, 2016. [Cited on page 24.]

M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. [Cited on pages xi, 62, 63, and 80.]

P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004. [Cited on pages 25 and 26.]

M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. [Cited on page 72.]

J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí. Yet another survey on image segmentation: Region and boundary information integration. In *European Conference on Computer Vision*, pages 408–422. Springer, 2002. [Cited on page 24.]

B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *International Conference on Computer Vision*, pages 670–677. IEEE, 2009. [Cited on page 25.]

R. Garnavi, M. Aldeen, M. E. Celebi, G. Varigos, and S. Finch. Border detection in dermoscopy images using hybrid thresholding on optimized color channels. *Computerized Medical Imaging and Graphics*, 35(2):105–115, 2011. [Cited on page 101.]

M. Godec, P. M. Roth, and H. Bischof. Hough-based tracking of non-rigid objects. *Computer Vision and Image Understanding*, 117(10):1245–1256, 2013. [Cited on pages 66 and 77.]

M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2141–2148. IEEE, 2010. [Cited on pages 66, 77, and 90.]

C. Gu, J. J. Lim, P. Arbeláez, and J. Malik. Recognition using regions. In *Conference on Computer Vision and Pattern Recognition*, pages 1030–1037. IEEE, 2009. [Cited on page 46.]

S. Gu, Y. Zheng, and C. Tomasi. Critical nets and beta-stable features for image matching. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010. [Cited on page 44.]

S. Hare, A. Saffari, and P. H. Torr. Struck: Structured output tracking with kernels. In *International Conference on Computer Vision*, pages 263–270. IEEE, 2011. [Cited on pages 66 and 80.]

C. G. Harris, M. Stephens, et al. A combined corner and edge detector. In *Alvey Vision Conference*, volume 15, pages 10–5244. Citeseer, 1988. [Cited on pages 18 and 19.]

H. Huang and P. Bergstresser. A new hybrid technique for dermatological image registration. In *BioInformatics and BioEngineering, 2007 IEEE International Conference on*, pages 1163–1167. IEEE, 2007. [Cited on page 102.]

D. E. Ilea and P. F. Whelan. Image segmentation based on the integration of colour-texture descriptors: A review. *Pattern Recognition*, 44(10-11):2479–2501, 2011. [Cited on page 24.]

Z. Kalal, J. Matas, and K. Mikolajczyk. Pn learning: Bootstrapping binary classifiers by structural constraints. In *CVPR, 2010 IEEE Conference on*, pages 49–56. IEEE, 2010. [Cited on pages 56, 63, 66, and 80.]

S. Keel, Z. Li, J. Scheetz, L. Robman, J. Phung, G. Makeyeva, K. Z. Aung, C. Liu, X. Yan, W. Meng, et al. Development and validation of a deep learning algorithm for the detection of neovascular age-related macular degeneration from color fundus photographs. *Clinical & Experimental Ophthalmology*, 2019. [Cited on page 4.]

P. Koniusz and K. Mikolajczyk. Segmentation based interest points and evaluation of unsupervised image segmentation methods. In *BMVC*, pages 1–11, 2009. [Cited on page 44.]

I. Kreso, S. Segvic, and J. Krapac. Ladder-style densenets for semantic segmentation of large natural images. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 238–245, 2017. [Cited on pages 120 and 128.]

M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernández, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder. The visual object tracking vot2015 challenge results. In *International Conference on Computer Vision Workshops*, pages 1–23, 2015. [Cited on page 79.]

B. Kumar, G. Carneiro, I. Reid, et al. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 5385–5394, 2016. [Cited on pages 18 and 22.]

J. Kwon and K. M. Lee. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*, pages 1208–1215. IEEE, 2009. [Cited on page 60.]

J. Kwon and K. M. Lee. Visual tracking decomposition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1269–1276. IEEE, 2010. [Cited on pages 56, 60, 66, and 79.]

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015. [Cited on page 3.]

Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *International Conference on Computer Vision*, pages 1995–2002. IEEE, 2011. [Cited on pages 66 and 77.]

S. Leutenegger, M. Chli, and R. Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *International Conference on Computer Vision*, pages 2548–2555. IEEE, 2011. [Cited on pages 17, 18, and 21.]

A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi. Turbopixels: Fast superpixels using geometric flows. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 31 (12):2290–2297, 2009. [Cited on pages 25 and 27.]

A. Levinshtein, C. Sminchisescu, and S. Dickinson. Multiscale symmetric part detection and grouping. *International journal of computer vision*, 104(2):117–134, 2013. [Cited on page 90.]

F. Li, T. Kim, A. Humayun, D. Tsai, and J. Rehg. Video segmentation by tracking many figure-ground segments. In *International Conference on Computer Vision*, pages 2192–2199, 2013. [Cited on pages vi, 64, 66, 67, 75, 76, and 77.]

Z. Li and J. Chen. Superpixel segmentation using linear spectral clustering. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1356–1363, 2015. [Cited on pages 25 and 28.]

J. Lim, D. A. Ross, R.-S. Lin, and M.-H. Yang. Incremental learning for visual tracking. In *Advances in neural information processing systems*, pages 793–800, 2004. [Cited on pages 63, 66, and 79.]

F. LIRIS. The visual object tracking vot2014 challenge results. 2014. [Cited on pages 63, 64, and 75.]

M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation. In *CVPR 2011*, pages 2097–2104. IEEE, 2011. [Cited on pages 25 and 27.]

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3730–3738, 2015. [Cited on page 4.]

D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157. Ieee, 1999. [Cited on pages 4, 5, 17, 57, and 93.]

D. G. Lowe. Distinctive image features from scale-invariant keypoints. volume 60, pages 91–110. Springer, 2004. [Cited on pages 17, 19, 35, 41, 42, 71, 88, 108, 117, 118, 125, and 126.]

L. Lu and G. D. Hager. A nonparametric treatment for location/segmentation based visual tracking. In *Computer Vision and Pattern Recognition (CVPR), 2007 IEEE Conference on*, pages 1–8. IEEE, 2007. [Cited on page 56.]

Z. Ma and J. M. R. Tavares. A novel approach to segment skin lesions in dermoscopic images based on a deformable model. *IEEE journal of biomedical and health informatics*, 20(2):615–623, 2015. [Cited on page 101.]

I. Maglogiannis. Automated segmentation and registration of dermatological images. *Journal of Mathematical Modelling and Algorithms*, 2(3):277–294, 2003. [Cited on page 102.]

E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger. Adaptive and generic corner detection based on the accelerated segment test. In *Computer Vision–ECCV 2010*, pages 183–196. Springer, 2010. [Cited on page 72.]

B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on circuits and systems for Video Technology*, 11(6):703–715, 2001. [Cited on page 5.]

M. Martín Redondo. Evaluación comparativa de técnicas de detección y descripción de putnos de interés en imágenes. Master's thesis, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 2016. [Cited on pages 7, 13, 29, 30, and 131.]

B. Mcgregor. Automatic registration of images of pigmented skin lesions. *Pattern Recognition*, 31(6): 805–817, 1998. [Cited on page 102.]

S. W. Menzies, A. Gutenev, M. Avramidis, A. Batrac, and W. H. McCarthy. Short-term digital surface microscopic monitoring of atypical or changing melanocytic lesions. *Archives of dermatology*, 137(12): 1583–1589, 2001. [Cited on page 110.]

K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. 2001. [Cited on page 19.]

K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European conference on computer vision*, pages 128–142. Springer, 2002. [Cited on page 19.]

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005. [Cited on pages 5, 20, 30, 45, 48, 49, and 57.]

K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International journal of computer vision*, 65 (1-2):43–72, 2005. [Cited on pages 5 and 19.]

F. Navarro. Local features and superpixel techniques. Technical report, Queen Mary University of London, 2014a. [Cited on pages 7 and 32.]

F. Navarro. Viability study of video applications for outdoor cameras - uav vision. Technical report, Universidad Autónoma de Madrid, 2014b. [Cited on page 7.]

F. Navarro and A. Cavallaro. Local descriptors and superpixels. Technical report, Queen Mary University of London, 2014. [Cited on page 29.]

F. Navarro, M. Escudero-Viñolo, and J. Bescós. Enhancing region-based object tracking with the sp-sift feature. In *In proc. of International Workshop on Content-Based Multimedia Indexing*, pages 1–4. IEEE, 2014a. [Cited on pages 6, 55, and 68.]

F. Navarro, M. Escudero-Viñolo, and J. Bescos. Sp-sift: enhancing sift discrimination via super-pixel-based foreground–background segregation. *IET Electronics Letters*, 50(4):272–274, 2014b. [Cited on pages 6, 41, 68, and 72.]

F. Navarro, M. Escudero-Viñolo, and J. Bescós. Accurate segmentation and registration of skin lesion images to evaluate lesion change. *IEEE Journal of Biomedical and Health Informatics*, 23(2):501–508, 2018a. [Cited on pages 6, 87, and 99.]

F. Navarro, M. Escudero-Viñolo, and J. Bescós. Hpstr: Homography point-based shape-fitted tracker. 2018b. [Cited on pages 6 and 55.]

P. Neubert and P. Protzel. Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In *International Conference on Pattern Recognition*, pages 996–1001. IEEE, 2014. [Cited on page 27.]

H. T. Nguyen and A. W. Smeulders. Fast occluded object tracking by a robust appearance filter. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 26(8):1099–1104, 2004. [Cited on pages 56 and 57.]

K. Nummiaro, E. Koller-Meier, and L. Van Gool. An adaptive color-based particle filter. *Image and vision computing*, 21(1):99–110, 2003. [Cited on pages 66 and 79.]

T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996. [Cited on page 46.]

Y. Ono, E. Trulls, P. Fua, and K. M. Yi. Lf-net: learning local features from images. In *Advances in Neural Information Processing Systems*, pages 6234–6244, 2018. [Cited on page 4.]

S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan. Locally orderless tracking. *International Journal of Computer Vision*, 111(2):213–228, 2015. [Cited on page 56.]

D. C. W. Pao, H. F. Li, and R. Jayakumar. Shapes recognition using the straight line hough transform: Theory and generalization. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11): 1076–1089, 1992. [Cited on page 105.]

L. Patino, T. Cane, A. Vallee, and J. Ferryman. Pets 2016: Dataset and challenge. In *Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016. [Cited on page 79.]

F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016. [Cited on pages vii, 94, and 95.]

D. A. Perednia and R. G. White. Automatic registration of multiple skin lesions by use of point pattern matching. *Computerized medical imaging and graphics*, 16(3):205–216, 1992. [Cited on page 102.]

X. Ren and J. Malik. Learning a classification model for segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2003 IEEE Conference on*, pages 10–17. IEEE, 2003. [Cited on page 56.]

X. Ren and J. Malik. Tracking as repeated figure/ground segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2007 IEEE Conference on*, pages 1–8. IEEE, 2007. [Cited on page 56.]

S. Rota Bulò, L. Porzi, and P. Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*, pages 5639–5647, 2018. [Cited on pages 120 and 128.]

E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *International Conference on Computer Vision*, pages 2564–2571. IEEE, 2011. [Cited on pages 17, 18, and 21.]

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. [Cited on page 4.]

P. Salembier and F. Marqués. Region-based representations of image and video: segmentation tools for multimedia services. *IEEE Transaction on Circuits and Systems for Video Technology*, 9(8):1147–1169, 1999. [Cited on page 24.]

J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof. Prost: Parallel robust online simple tracking. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 723–730. IEEE, 2010. [Cited on pages 56, 60, 66, and 79.]

C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 19(5):530–535, 1997. [Cited on page 18.]

L. Shafarenko, M. Petrou, and J. Kittler. Automatic watershed segmentation of randomly textured color images. *IEEE Transactions on Image Processing*, 6(11):1530–1544, 1997. [Cited on page 35.]

J. Shi and J. Malik. Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107, 2000. [Cited on pages 25 and 26.]

J. Shi and C. Tomasi. Good features to track. In *Computer Vision and Pattern Recognition (CVPR), 1994 IEEE Conference on*, pages 593–600, Jun 1994. doi: 10.1109/CVPR.1994.323794. [Cited on page 5.]

H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016. [Cited on page 4.]

G. Shu, A. Dehghan, and M. Shah. Improving an object detector and extracting regions using superpixels. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3721–3727, 2013. [Cited on page 25.]

M. Silveira, J. C. Nascimento, J. S. Marques, A. R. Marçal, T. Mendonça, S. Yamauchi, J. Maeda, and J. Rozeira. Comparison of segmentation methods for melanoma diagnosis in dermoscopy images. *IEEE Journal of Selected Topics in Signal Processing*, 3(1):35–45, 2009. [Cited on page 101.]

K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 36(8):1573–1585, 2014. [Cited on page 18.]

A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis & Machine Intelligencev*, 36(7): 1442–1468, 2013. [Cited on pages 56, 57, and 59.]

A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 36(7):1442–1468, 2014. [Cited on pages 62, 63, 64, 65, 66, 75, 79, and 80.]

D. Stutz, A. Hermans, and B. Leibe. Superpixels: An evaluation of the state-of-the-art. *Computer Vision and Image Understanding*, 166:1–27, 2018. [Cited on pages 25 and 29.]

D.-N. Ta, W.-C. Chen, N. Gelfand, and K. Pulli. Surftrac: Efficient tracking and continuous object recognition using local feature descriptors. In *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*, pages 2937–2944. IEEE, 2009. [Cited on page 56.]

F. Tiburzi, M. Escudero, J. Bescós, and J. M. Martínez. A ground truth for motion-based video-object segmentation. In *IInternational Conference on Image Processing*, pages 17–20. IEEE, 2008. [Cited on page 52.]

E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. In *Computer Vision and Pattern Recognition (CVPR), 2008 IEEE Conference on*, pages 1–8. IEEE, 2008. [Cited on page 57.]

E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 32(5):815–830, 2009. [Cited on pages 5, 21, and 45.]

D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg. Motion coherent tracking using multi-label mrf optimization. *International journal of computer vision*, 100(2):190–202, 2012. [Cited on page 66.]

H. Tsao, J. M. Olazagasti, K. M. Cordoro, J. D. Brewer, S. C. Taylor, J. S. Bordeaux, M.-M. Chren, A. J. Sober, C. Tegeler, R. Bhushan, et al. Early detection of melanoma: reviewing the abcdes. *Journal of the American Academy of Dermatology*, 72(4):717–723, 2015. [Cited on page 102.]

T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *International journal of computer vision*, 59(1):61–85, 2004. [Cited on page 44.]

T. Tuytelaars, K. Mikolajczyk, et al. Local invariant feature detectors: a survey. *Foundations and trends® in computer graphics and vision*, 3(3):177–280, 2008. [Cited on page 5.]

M. Van den Bergh, X. Boix, G. Roig, B. de Capitani, and L. Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *European conference on computer vision*, pages 13–26. Springer, 2012. [Cited on pages 25 and 28.]

A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *European conference on computer vision*, pages 705–718. Springer, 2008. [Cited on pages 25 and 27.]

Y. Verdie, K. Yi, P. Fua, and V. Lepetit. Tilde: a temporally invariant learned detector. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 5279–5288, 2015. [Cited on pages 18 and 22.]

L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):583–598, 1991. [Cited on pages 25 and 27.]

P. Viola, M. Jones, et al. Rapid object detection using a boosted cascade of simple features. volume 1, pages 511–518, 2001. [Cited on page 18.]

J. Wang and X. Wang. Vcells: Simple and efficient superpixels using edge-weighted centroidal voronoi tessellations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 34(6):1241–1247, 2012. [Cited on page 90.]

M. Wang, X. Liu, Y. Gao, X. Ma, and N. Q. Soomro. Superpixel segmentation: A benchmark. *Signal Processing: Image Communication*, 56:28–39, 2017. [Cited on pages 25 and 29.]

S. Wang, H. Lu, F. Yang, and M.-H. Yang. Superpixel tracking. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1323–1330. IEEE, 2011. [Cited on pages 25, 56, 58, 60, 66, and 80.]

W. Wang and J. Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5): 2368–2378, 2017. [Cited on pages 120 and 128.]

X. Wei, Q. Yang, Y. Gong, N. Ahuja, and M.-H. Yang. Superpixel hierarchy. *IEEE Transactions on Image Processing*, 27(10):4838–4849, 2018. [Cited on page 24.]

L. Wen, D. Du, Z. Lei, S. Z. Li, and M.-H. Yang. Jots: Joint online tracking and segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 2226–2234, 2015. [Cited on pages 64, 66, 67, 75, 77, and 80.]

W. Wu, A. Y. C. Chen, L. Zhao, and J. J. Corso. Brain tumor detection and segmentation in a crf (conditional random fields) framework with pixel-pairwise affinity and superpixel-level features. *International Journal of Computer Assisted Radiology and Surgery*, 9(2):241–253, Mar 2014. ISSN 1861-6429. doi: 10.1007/s11548-013-0922-7. URL https://doi.org/10.1007/s11548-013-0922-7. [Cited on page 25.]

Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2411–2418, 2013. [Cited on page 63.]

F. Yang, H. Lu, and M.-H. Yang. Robust superpixel tracking. *Image Processing, IEEE Transactions on*, 23(4):1639–1651, 2014. [Cited on pages xii, 66, 67, 79, 80, and 81.]

K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483. Springer, 2016. [Cited on pages 18 and 23.]

L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE transactions on medical imaging*, 36(4):994–1004, 2016. [Cited on page 101.]

X. Yuan, J. Yu, Z. Qin, and T. Wan. A sift-lbp image retrieval model based on bag of features. In *IEEE international conference on image processing*, pages 1061–1064, 2011. [Cited on page 46.]

D. Zhang, O. Javed, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 628–635, 2013. [Cited on page 66.]

Y. Zhang, R. Hartley, J. Mashford, and S. Burn. Superpixels via pseudo-boolean optimization. In *International Conference on Computer Vision*, pages 1387–1394. IEEE, 2011. [Cited on pages 25 and 26.]

Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial intelligence*, 78(1-2):87–119, 1995. [Cited on page 18.]

R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1265–1274, 2015. [Cited on pages 120 and 128.]

B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. [Cited on page 4.]

H. Zhou, X. Li, G. Schaefer, M. E. Celebi, and P. Miller. Mean shift based gradient vector flow for image segmentation. *Computer Vision and Image Understanding*, 117(9):1004–1016, 2013. [Cited on page 101.]

Y. Zhou, L. Ju, and S. Wang. Multiscale superpixels and supervoxels based on hierarchical edge-weighted centroidal voronoi tessellation. *IEEE Transactions on Image Processing*, 24(11):3834–3845, 2015. [Cited on page 90.]